The Impact of Incentive Payments on Outcomes for Common Experimental Methods

Stephanie Rosch (USDA ERS), Jacob Fooks (Cygna), Daniel Hellerstein (USDA ERS), Lori Lynch (University of Maryland), Kent Messer (University of Delaware), Gianna Short (USDA ERS), Sharon Raszap Skorbiansky (USDA ERS), Steven Wallander (USDA ERS), Collin Weigel (Johns Hopkins University), Eliana Zeballos (USDA ERS)

September 2019

Abstract:

According to the foundational rules set forth by Vernon Smith (1982), economics experiments need to provide payments to motivate subjects to care about their performance in the experiment. The payments must be salient to subjects and dominate any other considerations that could affect their performance (e.g., a desire to please the experimenter, innate enjoyment of the experiment activities, overcoming boredom). Improperly calibrated incentives increase the noisiness of subject responses, reducing the power of the experiment to detect treatment effects. However, what exactly constitutes the most appropriate calibration of incentive payment for a given experiment can be difficult to judge. We review the literature on the effects of incentive payments on outcomes in economics experiments involving real-effort tasks, auctions, multi-player games, market experiments, and risk elicitations. We also re-analyze published results to look for evidence that stronger incentives reduce the noisiness of experimental outcomes, consistent with a hypothesis that incentives improve the statistical power of experimental design.

For real effort tasks, we observe a pattern of decreased noisiness of outcomes at higher levels of piece-rate payments, quota payments, and tournament payments. We also observe patterns of decreased noisiness of outcomes at higher levels of incentives for conservation auctions, and two out of three types of risk elicitation tasks. However, the total number of studies available for these analyses and the range of payments used in those studies are too limited to present strong evidence of a relationship between incentives and precision of measured outcomes for these methods. Sufficient data was not available to analyze the results for multi-player games or market experiments, although the few available studies suggest potential for a relationship between incentives and outcomes for these methods. We are unable to rule out a non-monotonic relationship between payments and performance for any method reviewed and are unable to determine whether a minimum payment threshold exists that is common to all methods reviewed.

Keywords: experimental economics, incentives, statistical power

1.0 Introduction

As per the terms and conditions of the approved Information Collection Request 2016-12-0536-001 (OMB control number 0536-0076), USDA Economics Research Service (ERS) is required to submit "a literature review of the relationship between payment amount and performance [in experiments] that will provide a comprehensive picture of the research findings in the social and behavioral science research literature." In this white paper, we present a review of the literature on how payments affect participants' performance in economics experiments, and in other social and behavioral science disciplines that rely on similar methodologies as economic experiments.

There is a large literature dedicated to measuring the effect of payments on performance. Because of time and resource constraints, we restrict our review to a subset of methodologies that are particularly relevant to ERS research: real-effort tasks, conservation auctions, market experiments, and risk elicitations. We also provide a limited discussion of the literature on the effects of incentives in consumer auctions, single- and multi-player games, and choice experiments.

Our review included studies that found positive relationships between experimental payments and performance, negative relationships, null relationships, and non-linear relationships. However, these studies differed from each other along many critical dimensions such as methodology used, population sampled, type of outcome measured, and statistical power of the original experimental design. In order to summarize these results across disparate experiments, we discuss the literature separately for each methodology and reanalyze the results of each study using a measure of performance that is consistent with economic theory and statistical calculations of experimental power.

We conduct our review in the light of three key questions relevant for USDA ERS experimental research policy:

1) Do payments decrease the noisiness of experimental outcomes?
2) Is the effect of payments on the noisiness of experimental outcomes monotonic over the range of payments sampled?
3) What is the level of payments necessary to decrease the noisiness of experimental outcomes to meet the criteria of an adequately powered statistical data collection?

Overall, we find either a weak negative or null relationship between payments and noisiness of experimental outcomes for the methods analyzed. However, our ability to characterize these relationships is severely constrained by the limited number of studies conducted and a lack of studies using large incentives. The bulk of studies reviewed used a narrow range of relatively small incentives. We are unable to rule out a non-monotonic relationship between payments and performance for any method reviewed and are unable to determine whether a minimum payment threshold exists that is common to all methods reviewed. Our findings suggest that pre-testing is likely to be necessary to appropriately calibrate incentive payments for a given experimental design.

Section 2 provides background information on the use of incentives in economics and psychology experiments, roles of participation and incentive payments in experiments, and the distinction between hypothetical bias and incentive calibration. Section 3 presents our methodology for cataloging the literature and estimating relationships between experimental outcomes and strength of incentives, including a conceptual model of how incentives and subjects' effort create outcomes in economic

experiments. Section 4 examines the effects of performance incentives on real-effort tasks. Section 5 provides results for other experimental methods relevant for ERS, including auctions, multi-player games, market experiments, and risk elicitations. In section 6, we synthesize the results of our review and identify critical gaps in the literature to address with future research.

2.0 Background

2.1 Incentives in Economics and Psychology

Performance payments are a hallmark of economics experiments and one of the key ways in which experiment methodologies used in economics differ from those used in psychology (Hertwig and Ortmann, 2001). The difference stems from the two fields' views of incentive theories. Economists hypothesize that incentives are necessary to induce individuals to undertake a costly effort. Similarly, to encourage "better" work – for participants to pay greater attention to a task – the payments provided in the experiment must be "better" than the benefits of doing other tasks not related to the experiment. Psychologists, on the other hand, generally hypothesize that individuals reveal truthful behaviors in experiments because of intrinsic motivations such as a desire to please, innate enjoyment of the activity, and/or a desire to overcome boredom (Camerer and Hogarth, 1999). Because of such intrinsic motivations, subjects should be willing to perform tasks without receiving an external incentive. Furthermore, both psychology and economics researchers have hypothesized that, in some cases, performance incentives conflict with intrinsic motives, leading to a crowding-out effect that reduces or eliminates pro-social behaviors (Gneezy et al., 2011).

Outcomes of economic experiments - such as willingness to pay in auctions, the amount of money contributed to a public good, the amount of profit earned in a market, and choices made in risk-elicitation experiments - are statistical outcomes, and as such, are measured with noise. This error is always assumed to have a random statistical noise component, but may also have a non-random component caused by human error. For example, human error could result in systematically over bidding in an auction relative to the value of the good being sold, generating an estimated willingness to pay for the good that over-estimates the individual's true willingness to pay.

If the economic theory of incentives is correct, adequate performance payments induce subjects to work harder at the experimental tasks, which reduces the noisiness of individuals' actions and improves the power of an experiment to detect treatment effects. The theory also implies that performance incentives smaller than the cost of that effort should have no impact on variance in the subjects' responses.

These predictions have been upheld in the early literature reviewing both economic and psychology experiments. In a review of 31 early experimental studies, Smith and Walker (1993) concluded that "in virtually all cases, rewards reduce the variance of the data around the predicted outcome" but do not affect the mean outcome in a significant way. Camerer and Hogarth's review of 74 economic and psychology experiments found more variation in the effect of performance incentives on outcomes. Some studies showed that a larger incentive improved mean performance, but most found that the size of an incentive did not affect mean performance in the experiment. In some cases, a greater incentive did not change the mean difference in the outcome variable but did reduce the variance of performance. The authors concluded that the size of the incentive had little impact when (1) the return

to greater effort was small, (2) it was difficult for subjects to determine how to do better, and (3) the task was too easy.

To be effective, incentives need to be calibrated to the cost of the effort required in the experimental design. But the literature has not provided much guidance on how to tell if the chosen level of incentive payments is sufficient for the given experimental design.

Vernon Smith, who received the Nobel Prize in Economic Sciences in 2002 for his work foundational work on experimental economics, identified two requirements for performance payments in economics experiments (Smith, 1982). First, the payments must be salient – they must be provided in a form that is meaningful to the participants[1] (e.g., subjects in the United States should not be paid in Italian Lira) and must increase in value with the quality of the outcome. Second, the payment amounts must be dominant – the amounts must provide stronger motivation than any of the participants' intrinsic (and often unobservable) motivations. Unfortunately, the literature provides little guidance regarding how to calibrate such payments for specific experiment tasks and participant populations, and the few studies that have addressed calibration have rarely described the procedures used to determine such values. Instead experimental economists have relied upon general rules of thumb when setting the incentives levels, such as paying an hourly wage roughly equivalent to the expected wage of the sample population.

While incentive payments need to be large enough to compensate for the effort exerted in the experiment, this does not necessarily imply that only very large incentive payments will be adequate to achieve payment dominance. Camerer and Hogarth (1999) found that going from no incentive payments (hypothetical) to low incentive payments had a greater effect on outcomes than going from a low to high payments incentive payments, a finding consistent with diminishing marginal returns to payments. Camerer and Hogarth also found that the effects of incentive payments were methodology specific, implying that an incentive payment adequate to achieve dominance in one type of experimental task may not be sufficient to achieve dominance in a different type of task.

2.2. Participation Fees vs Incentive Payments

Experimentalists use two types of fees: participation fees and performance incentives. Participation fees impact the types of subjects recruited for the experiment, while performance payments provide incentives to exert effort to complete the experimental task. Use of participation fees is common for other types of data collections such as focus groups or surveys. Use of performance payments is uncommon outside of experiments.

Participation fees are widely used in economic experiments and were provided in the majority of studies reviewed for this report. However, participation fees are not universally used. Some experimental

[1] Economic experiments typically use cash payments instead of other incentives (i.e. small gifts such as hats or mugs, or large items such as trips to exotic locations). This is because the value to purchase goods and services with the cash is likely to be the same across subjects, while the consumption value of the object may differ across subjects. Non-monetary incentives can also motivate performance in laboratory experiments (see Cassar and Meier, 2018). Non-monetary incentives have been shown to motivate effort better than low levels of incentive payments, but literature is inconclusive about whether these factors are complements or substitutes for incentive payments.

methods, such as auctions and incentivized choice experiments, tend to provide a single payment instead of providing a separate incentive payment and participation fee. This is done so that subjects will treat both payments as a single endowment to be used to pay for the good being auctioned or purchased. Classroom experiments rarely include participation fees or incentive payments as subjects are recruited from class participants and course credit is typically offered as the incentive for exerting effort in the experiment.

Increased recruitment incentive payments have been shown to increase willingness to participate in surveys and experiments (Korn and Hogan 1992; Bently and Thacker 2004; Buhrmester, Kwang, and Gosling 2011). One argument for having high participation fees is that you can recruit a more demographically representative pool of participants, especially in terms of income. For example, Slonima, et al. (2013) found that people who elected to participate in their lab experiment had significantly less income and more leisure time than non-participants. However, participation fees do have the potential to introduce bias to the sample. Harrison, Lau, and Rustrom (2009) found that the use of predetermined, guaranteed participation fees resulted in a more risk-averse sample than would otherwise have been the case.

Participation fee amounts can range substantially depending on the experiment setting and recruitment style. Lab-based experiments require schedule alignment and travel in addition to the effort required for the actual participation. Participation fees for in-lab experiments can range from minimum wage equivalents to upwards of $30-$40 for an hour of time. In contrast, online experiments can be completed on the participant's schedule and do not require travel. Short, online surveys on platforms such as Amazon's MTurk often have participation fees of a few cents. While ethics guidelines for online surveys encourage offering minimum wage equivalents as participation fees, studies have found that subjects on MTurk to have a reservation wage of only $1.38 per hour and an average effective hourly wage of $4.80 (Mason and Suri 2012).

2.3 Hypothetical Bias vs Uncalibrated Incentives

There are two issues for incentive payments: (1) whether they are needed at all, and if they are needed, (2) what level of payment are necessary to properly incentivize subjects in the experiment. The first issue is referred to as hypothetical bias, and its existence has been well documented in the literature for eliciting willingness-to-pay (WTP). The second issue has received less attention from the literature and is the main focus of this report.

WTP values from hypothetical elicitation tasks have been found to nearly always exceed WTP values for the same goods from non-hypothetical elicitation tasks (Chang et al., 2009; List and Gallet, 2001; Little and Berrens, 2004; Murphy et al., 2005). The rationale used to explain these differences is that subjects do not put as much cognitive effort into hypothetical choices and do not have an incentive to reveal their true values (i.e., the choices are not incentive-compatible). Gracia et al. (2011) conducted experiments to compare WTP estimates from non-hypothetical choice experiments and experimental auctions and found that the valuations often differed. Johansson-Stenman and Svedsäter (2008) set up a choice experiment eliciting environmental values with actual and hypothetical trade-offs and found that the between-subject design suffered from greater hypothetical bias than the within-subject design using otherwise identical scenarios. Yue and Tong (2009) combined a hypothetical experiment with a non-hypothetical choice mechanism to investigate consumers' WTP for organic, local, and organic plus local

attributes of fresh produce. They found only limited hypothetical bias when real products were used in the hypothetical experiment. Fifer et al. (2014) explored responses of motorists in stated-choice experiments to study hypothetical bias and the extent it could be mitigated using cheap talk and certainty scales. They found that the estimates from the stated-choice model were prone to hypothetical bias and that the mitigation techniques could potentially compensate for it. Chang et al. (2009) compared the ability of hypothetical choices, non-hypothetical choices, and non-hypothetical rankings and three discrete-choice econometric models to predict actual retail shopping behavior in three product categories (ground beef, wheat flour, and dishwashing liquid). Overall, they found that their estimates had a high level of external validity and their results suggested that non-hypothetical choices and non-hypothetical rankings in particular outperformed the hypothetical choice experiment in predicting retail sales (Chang et al., 2009).

3.0 Methodology

In this section, we review the implications of statistical power for testing the predictions of economic theory of incentives. Then we explain our methodological approach for analyzing the effect of incentives on statistical outcomes, describe our criteria for including and excluding studies, and describe the process used to peer-review our methods.

3.1 Incentives, Outcomes, and Statistical Power

A fundamental principal in economic theory is that people can be motivated by incentives. Economic experiments generally require subjects to complete a difficult task – cognitively difficult, physically difficult, or both – in which the experiment's outcomes depend on subjects' aggregate performance of the task. According to incentive theory, people work harder to achieve a difficult task when the personal reward is greater.

This connection between personal rewards, effort exerted in a task, and effort-dependent experiment outcomes has two ramifications for the statistics of the experiment outcomes. First, the mean outcome across all subjects can increase with increasing subject effort. Second, the variance in individuals' outcomes can decrease with increasing subject effort. To see why these relationships occur, consider a hypothetical experiment that measures how fast a random group of 500 individuals can run 100 meters. Because individuals differ in their ability to run fast and motivation to run 100 meters, the experiment generates a distribution of completion times across those random 500 individuals. If each individual could receive a "meaningfully large" financial incentive for completing the 100 meters in 30 seconds or less, the distribution of completion times would have a lower mean and lower variance than a distribution with no incentive. The higher mean occurs because individuals who would otherwise take more than 30 seconds to run 100 meters are motivated to run faster to earn the financial prize. The decrease in variance occurs because of physical limits on individuals' abilities to run fast. Individuals who already ran 100 meters in less than 30 seconds cannot improve their times as much as individuals who took longer than 30 seconds can, thus compressing the variance of the distribution of individual completion times.

To see why experimentalists might want to incentivize the effort exerted by individuals in an experiment, assume that a hypothetical experiment measuring 500 individuals completing a 100-meter

dash was measuring the effect of different types of shoe soles on individuals' race times. The individual outcome measured in the experiment is the time each person requires to complete a single 100-meter dash. The randomly assigned treatment is the type of shoe worn in the race, and the statistic of interest for the experiment is the average difference in race times across types of shoes. Without an incentive, it would be difficult to detect an effect of each type of sole from the average running time for all 500 individuals given natural variations in their effort based on their intrinsic motivations. With a "meaningfully large" financial incentive, however, intrinsic motivations would be overwhelmed by the desire to be financially rewarded, and the precision of the average times measured for each shoe type would improve, making it easier to detect a treatment effect from shoe type.

To formalize this analysis, consider a hypothesis test for unequal means in the treated and control groups. The sample size required for the treated and control groups is calculated based on the desired effect size and tolerance for statistical error:

$$n_{i,j} = 2\left(\frac{Z_{1-\alpha/2} + Z_{1-\beta}}{ES}\right)^2 = 2\sigma^2\left(\frac{Z_{1-\alpha/2} + Z_{1-\beta}}{\left|\mu_i - \mu_j\right|}\right)^2$$

Here *i* and *j* are the treatment and control groups, $n_{i,j}$ are the number of individuals to sample for the treatment and control group, $Z_{1-\alpha/2}$ and $Z_{1-\beta}$ are the critical values from a standard normal distribution for the chosen levels of Type I and Type II errors, *ES* is the desired effect size, μ_i and μ_j are the population means for the treatment and control groups, and σ is the standard deviation of the outcome conditional on the incentive payment level[2]. Under the economic theory of incentives, this equation becomes:

$$n_{i,j} = 2\left[\frac{Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}}{ES(\pi)}\right]^2 = 2\left[\sigma(\pi)\right]^2\left[\frac{Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}}{\left|\mu_i(\pi) - \mu_j(\pi)\right|}\right]^2$$

The effect size is a function of the chosen incentive level, π. Incentives which increase subjects' effort will decrease the variance of outcomes and increase the difference in mean outcomes between the treatment and control groups. The combined effect of these changes will increase the desired effect size, and allowing the outcomes to be measured within the tolerance for statistical error at a lower sample size than would be required without incentives.

Because power is proportional to the square of the effect size, small changes in the effect size can have sizable implications for the number of subjects required in an experiment. For example, consider a hypothetical experiment with two treatments (one treatment, and one control) with 100 subjects per treatment and no incentive payments. If increasing incentives yielded a 10 percent increase in effect size, then that experiment would be adequately powered to detect a treatment effect with 83 subjects per treatment, potentially allowing the experiment to be conducted as a lower total cost than without the incentives.

However, increasing the size of an incentive may not necessarily lead to more-accurate experiment outcomes in all cases. If an incentive is already inducing subjects to exert their maximum effort,

[2] The standard deviation uses the pooled variance for the treated and untreated populations.

increasing the incentive will not lead to greater effort or result in any increase in the power of the experiment. If subjects are not exerting maximum effort and the incentive provided is not meaningfully large, subjects may not respond to it (e.g. Cadsby et al. (2007), Georgellis et al. (2010), Ashraf et al. (2014), and Levitt et al. (2016)) . In that case, the measured outcomes will remain dependent on subjects' intrinsic motivations to perform the task, which is likely hard to observe. In the hypothetical race experiment, for example, raising the incentive from one cent to ten cents, a ten-fold increase, likely would not motivate subjects to run faster. Subjects might even run more slowly than in a no-incentive case because they felt disincentivized by the low payment rate (e.g. Bowles (2008), Mellstrom and Johannesson (2008), Gneezy et al. (2011), and Rode et al. (2015)). Furthermore, subjects may consider an incentive to be meaningfully large for one type of task but not for another (see, for example, Bonner et al. (2000)). Thus the incentives offered need to be calibrated for the task and subject pool in order to be sure that they are having the desired effect on statistical power.

3.2 Calibrating Incentive Payments

Accurate calibration of the level of incentive provided depends on the relationships between the incentive, the degree of effort likely to be exerted, and the precision of measures of the outcomes in a given experiment design. Poorly-calibrated incentives should have either no effect on statistical power or act to decrease the statistical power to detect treatment effects and increase the noisiness of estimated outcomes such as risk attitude and willingness to pay. It could also bias estimates of aggregate statistics such as the market-clearing price and average contribution in a public good game. Well-calibrated incentives should increase the statistical power to measure outcomes and treatment effects.

A *statistically efficient calibration* of incentives balances the likelihood of generating false positive (type I) and false negative (type II) errors based on the desired number of subjects and treatment effect size. An *economically efficient calibration* would balance the marginal cost of effort with the marginal improvement in precision of outcomes subject to the budget for incentive payments. A calibration that combines both statistical and economic efficiency would choose incentives that balance the marginal cost of effort with the marginal improvement in precision of estimated outcomes subject to the tolerance for type I and type II errors and the budget constraint.

The efficient level of incentive to use depends on how the effect size varies with payment levels for a given task, experimental environment, and subject pool. For example, running one mile is more physically demanding than running one hundred meters, thus the efficient incentive payment for the more challenging task will be greater than for the less challenging task. Similarly, running one hundred meters on a slippery track is more challenging than running when the track surface has good traction, and thus the efficient incentive payments could vary with track condition. Also, the efficient level of incentives for subjects who like running will be lower than the efficient level of incentives for subjects who abhor running.

Experimentalists may need to pre-test different incentive schemes before conducting the experiment in order to verify that their chosen incentive scheme is adequate to meet the desired tolerance for type I and type II error. Ex-post outcomes from the experiment are not sufficient to demonstrate that the incentives were well-calibrated. An experiment with improperly calibrated incentive payments may still generate precisely measured outcomes if there is little variation in outcomes across subjects regardless of incentives.

3.3 Testable predictions from Incentive Theory and Empirical Approach

The economic theory of incentives provides several testable predictions. For a given experimental task, environment, and subject pool:

1) There will be a minimum threshold for an incentive payment such that payment below that threshold will have no effect on the precision of the experiment outcomes because it will fail to induce subjects to exert effort and therefore have no impact on the mean and variance of subjects' responses. This threshold will be higher for more cognitively and/or physically demanding tasks than for less demanding tasks.

2) There will be a maximum threshold for an incentive payment such that payment above that threshold will have no additional effect on the precision of the experiment outcomes because subjects will have already been willing to invest maximum effort and cannot increase their efforts further. This threshold will be higher for subjects who have greater aptitudes for the experimental task than for subjects with limited aptitude for the experimental task.

3) Between the minimum and maximum thresholds, there will be an increasing relationship between the magnitude of the incentive payment and the precision of the experiment outcomes. Subjects will respond to an increasing incentive by exerting an increasing amount of effort, which will increase the mean outcomes and reduce the amount of variance of the outcomes overall.

4) The minimum threshold for an incentive payment will increase with the income/wealth of the individual participant. Subjects with more income or wealth will be less motivated by low incentives than subjects with less income or wealth. Similarly, subjects who live in areas where the purchasing power of money is low will be less motivated by low incentives than subjects who live in areas where the purchasing power of money is high.

We use these predictions to guide our analysis of the literature. For each study reviewed, we graph the strength of the incentive used and the realized precision of the experimental outcome. Each graph shows the general relationship of incentives and precision of estimated outcomes, unconditional on other factors relevant for statistical power (e.g. number of subjects, number of treatments). For some methods, the data was not sufficient to analyze the studies graphically. For those methods, we characterize the evidence available as described in the relevant section.

We use the coefficient of variation (CV) of outcomes as our measure of precision of the experimental outcome. The CV scales each standard error by the sample mean, providing an intuitive way to compare variability across experiments that differ in the size of the mean effects. We calculate CVs for each study. For experiments with multiple treatments, we calculate CVs for each treatment within the study.

We calculate the strength of the incentive as the difference in earnings the subject would receive if he exerted no effort and the earnings the subject would receive if he exerted optimal effort. For simplicity, use the benchmark of a risk neutral, expected utility maximizer playing Nash equilibrium strategies for determining the optimal effort for each methodology. For studies that use multiple incentive schemes, we calculate the power of incentives separately for each incentive scheme. Hypothetical payments were treated as having zero incentive strength. For methods where we could not establish a zero effort benchmark, we use the average payment received in lieu of a measure of return to effort.

We convert the strength of incentive payments into dollar terms to be consistent across studies. Payments used in the studies reviewed included endowments given to subjects at the start of the experiment, per-unit payments, dollar values for profits earned in the experiment, and measures of the degree of risky payments made. Some real payments were converted from experimental currency into actual currency and/or paid out randomly by a lottery. In those cases, we calculate the strength of the incentives for the actual currency and/or expected lottery earnings rather than the experimental currency units.

We focus on experiment methodologies of relevance for USDA ERS experimental research: real effort tasks, auctions, multi-player games, market experiments, and risk elicitations. We analyze each methodology separately as the physical and cognitive burdens vary across methodologies. However, our review of the literature did not identify a means of ranking the physical and cognitive burdens among these tasks so we are unable to test if the minimum threshold for an incentive payment increases with the physical and cognitive burden of the task.

Our review of the literature also did not identify a standard method of measuring aptitude for different experimental methodologies in different populations. Therefore we are also unable to test if the maximum threshold for the incentive payment increases with subjects' aptitude. However, Dohmen and Falk (2011) found that the average time subjects spent on multiplying numbers (a type of real effort task) increased with the difficulty of the multiplication problem, that subjects tended to make more errors solving multiplication problems under higher strength incentive schemes, and that subjects with less aptitude for correctly multiplying numbers tended to sort themselves into a lower strength incentive scheme. These results are consistent with the existence of an average performance response to incentive strength that varies based on subject aptitude and cognitive burden of the task. However the range of incentive schemes used in the experiment was not sufficient to observe either a lower or an upper threshold for incentive effects on performance in this task.

To control for differences in the value of money across subject pools, we restrict our analysis to studies conducted in developed countries where $1 USD would have relatively similar levels of purchasing power. We included studies that sampled non-students and student populations because experiments with non-students often use the same range of payments as experiments using student subjects. However, we were unable to control for the income or wealth of subjects as this information was not reported for the majority of studies reviewed.

3.4 Criteria for Including Studies

To provide a comprehensive review of the experimental economics literature, we established inclusion and exclusion criteria to identify studies that were well designed and likely to have adequate power to detect treatment effects. We also prioritized studies that had within-subject variation in the payments. Table 1 presents those criteria.

Table 1: Study Inclusion and Exclusion Criteria

	Must Include	**Discretion to Include**	**Exclude**

Literature	Journals in economics and agricultural economics	Journals in management science, finance, psychology	None
Publication Status	Peer-reviewed journals (top and middle tier)	Peer-reviewed journals (lower tier), unranked journals, and unpublished manuscripts	Non-peer-reviewed journals
Study Location	United States, Canada, Europe, and Australia	Other countries	None
Sample Size	20 or more subjects per treatment	Less than 20 subjects per treatment	None

The criteria were chosen to favor rigorously peer-reviewed, adequately powered experimental studies involving subjects drawn from culturally similar populations (such as the United States, Canada, Europe, and Australia). We primarily included studies published in economic and agricultural economic journals but also opted to include some studies published in related fields based on their significance for the overall literature and design quality. The tier assigned to each journal was determined using ERS's journal-quality ranking methodology and was based on impact factors, H-index scores, and eigenfactors as reported by Research Papers in Economics (RePEc), Scimago, and Thomson-Reuters.[3] Studies published in the top and middle tiers were included, as were some peer-reviewed studies from lower-tier and unranked journals and unpublished manuscripts that provided particularly useful results. Studies published in journals that lacked peer review (including book chapters[4]) were excluded.

In terms of region, we prioritized studies conducted in the United States, Canada, Europe, and Australia because they are developed countries and the U.S. dollar has a similar degree of purchasing power in all of those regions. Studies conducted in developing countries would have provided a wider range of types of payments but also would have included populations with different social and economic valuations of money. A few particularly useful studies from those areas were included.

We screened studies based on the sample size per treatment as a proxy for adequately-powered studies. The National Agricultural Statistics Service (NASS) requires a minimum of 19 observations to report a statistic. Following that criterion, we set a threshold minimum of 20 subjects per treatment as an adequately powered study, and included studies with less than 20 subjects per treatment on a discretionary basis. In some of the experiment designs, the threshold of 20 subjects per treatment was not sufficient to consider the study as adequately powered[5], and we have noted those cases in the discussion.

We did not establish screening criteria for citation counts or publication dates because we preferred to infer each study's quality based on journal rankings and the study's design criteria to include high-quality, particularly informative studies in the review.

[3] Additional information about the ERS journal-ranking methodology is available upon request.
[4] Book chapters that were produced under a rigorous peer review process were eligible for inclusion.
[5] We originally excluded studies with fewer than 15 subjects per treatment, but in response to feedback received from the academic community.

3.5 Peer Review Process

We verified the methodological approach used to analyze experimental outcomes in presentations to audiences of experimental economists at:

- Economic Sciences Association North American Meeting in Richmond, VA; October 19-21, 2017
- Southern Economic Association (SEA) Meeting in Washington, DC; November 18-20, 2018.

Comments received from the audiences addressed the scope of the analyses, study inclusion/exclusion criteria, and the methodological approach. We adopted all of the audience recommendations except for one: a recommendation to reinforce our graphical analysis with a multivariate analysis that pooled all study results together. While we agree this would be a useful exercise, we were unable to complete this analysis due to time and resource constraints.

4.0 Effects of Performance Payments on Real-task and Real-effort Experiments

While many economic experiments use stylized games to mimic different types of real-world behaviors, there is a large literature on experiments involving real effort tasks. In those experiments, researchers ask participants to perform a real task that imposes an appreciable cost in terms of their time, attention, and effort and then measure the effect of an incentive on their performance.

Real effort tasks in the laboratory can closely mimic transactions for real effort in the workforce. In fact, when possible, researchers have worked with firms to run controlled experiments (Shearer 2004; Fehr and Goette 2007) or used firm data to measure the effect of changing incentives to workers' effort or productivity (Lazear 2000; Hamilton, Nickerson, and Owan 2003; Hubbard 2003; Zivin, Kahn, and Neidell 2019). Real effort experiments (or data from the real world) can illuminate the optimal payment scheme to attract pickers during harvest season, the necessary scheme and payment level to achieve higher output, and the necessary scheme and payment level to achieve higher quality output. However, ability to collect data from firms and use field experiments can be limiting, despite the importance of understanding incentive schemes to the agricultural community and USDA policy-making.

As an example, a recent study by Zivin, Kahn, and Neidell (2019) analyzes grape and blueberry farm worker productivity, who are compensated differently for their efforts. Both workers are paid a fixed wage up to a target threshold, and then a piece-rate. However, blueberry pickers are also paid a bonus for reaching the set target. The authors find that the bonus increases the likelihood of pickers to reach the target threshold by 15 percent over the first 10 days on the farm, while there is little improvements in reaching the threshold over the same time for grape pickers who are not paid a bonus. Studies such as these raise questions such as: (a) how would effort change given lower/higher piece-rates? (b) how would effort change given a lower/higher bonuses? However, field experiments are typically expensive and challenging to organize in large scales to test the sensitivity of payment schemes.

In the past, USDA ERS experiments have typically not involved real effort tasks though the social science literature has widely used real effort task experiments to gain understanding on how to better incentivize work. A large portion of the literature has concentrated on comparing different type of wage structures, such as fixed wage contracts to a piece-rate contracts (Lazear 2000; Paarsch and Shearer

2000; Burchett and Willoughby 2004; Dohmen and Falk 2011). These studies have concluded that piece-rate contracts are generally more efficient than fixed wages (Paarsch and Shearer 2000; Shearer 2004; Dohmen and Falk 2011). For example, Lazear (2000) uses data from a large manufacturing company that changed compensation schemes and finds productivity increased between 20 to 36 percent.

Starting with the hypothesis that performance pay is better at inducing higher productivity than fixed wages, the economic literature has also often used real effort tasks to test theory and justify the need for incentive payments in economic experiments (e.g. (Camerer et al. 1999; Croson 2005; Bardsley 2010)). We provide an overview of this literature to demonstrate the general relationship between incentive payments and outcomes of economic experiments in real effort tasks.

4.1. How real effort experiments work

The question of whether financial incentives can effectively motivate workers has received significant attention, not only from economists but from researchers in management, accounting, and psychology (Camerer et al. 1999; Bonner et al. 2000). Studies have used various types of real effort tasks when studying pay-for-performance incentives, such as:

- Vigilance/detection: Subjects respond to a stimulus (such as a light) with an action (such as pressing a button). This type of real effort experiment was often used in the psychology literature in the 1950s and 1960s.
- Memory: Subjects memorize information during the experiment, such as words, sentences, or numbers. This type of real effort experiment was often used in the psychology literature in the 1960s.
- Judgment: Subjects guess or evaluate something, such as the number of coins in a jar. This type of real effort experiment was often used in psychology experiments in the 1980s and 1990s.
- Production and clerical: Subjects perform simple tasks such as sorting parts, constructing a model, and stuffing envelopes. This type of real effort experiment most closely resembles any physical aspect of real-world work. For example, constructing a model can resemble assembly-line work. Stuffing envelopes or inputting data can resemble clerical jobs.
- Problem-solving/reason: Subjects attempt to achieve a goal, such as solving a puzzle or a mathematical equation. This type of real effort experiment most closely resembles any mental aspect of real-world work. Subjects are asked to solve problems, easy or complex that they may not have previously encountered.

Laboratory real effort tasks and field experiments are similar, in the sense that they both require the participants to expend true effort in accomplishing the experiment's outcome. However, field experiment tasks are performed in the natural environment of the participants. Some early real effort psychology experiments toed the line by "hiring" participants to jobs they believed to be real (e.g. Pritchard et al. 1976). Economic field experiments testing the sensitivity of payment schemes involve a wide variety of activities, including planting trees (Shearer 2004), delivering messages by bicycle (Fehr and Goette 2007), or harvesting fruit (Bandiera 2007).

4.2 Norms for Payments Used in Real-task and Real-effort Experiments

Typically, real effort experiments pay low rates per tasks. The range in pay differs widely by the type of pay-for performance scheme. Experimenters commonly use one of two pay-for-performance schemes in economic studies:

- Piece-rates: Subjects receive a predetermined amount of money linked to their performance or output, such as $1.00 per task completed. The majority of studies range between 5 and 50 cents.
- Quotas: Subjects earn a flat rate until they reach a performance threshold and then either receive a bonus or earn piece-rate payments for performance or output beyond the threshold.
- Tournament: Subjects are paid for their relative performance. Subjects are ranked and given a price for placing in the top spots. This scheme is similar to a quota; in a quota subjects "win" by reaching the threshold, while in a tournament subjects "win" by reaching the top tier.

An additional pay-for-performance schemes found in the literature on experimental psychology is variable ratio, rewarding some of the units of output. Because this scheme is not fully tied to performance, we do not address experiments that use it. Additionally, we do not examine studies which evaluate performance collectively based on the effort put forth by a group of subjects.

The reviewed studies also vary in how the control treatments are structured. For example, some piece-rate studies compare the effects of high piece-rates to low piece-rates, with the low piece-rate payment serving as the control treatment. In others, the control treatment is a flat or fixed payment equivalent to a salary and unrelated to performance. In a few cases, the control group receives no monetary compensation. Varying combinations of pay-for-performance schemes and bases answer related but different questions. For example, a base of a low piece-rate payment versus a high-piece-rate treatment provides information about the effect on effort of larger payments, while comparisons to a flat rate provides information regarding the effect of incentivizing subjects (with a specific rate) versus paying them a "salaried wage".

We do not examine the gains from switching from one type of payment scheme to the next. However, if a study compares the outcomes achieved under a quota system versus the outcome in a piece-rate, given that sufficient information is provided (mean and standard deviation) we record the results.

4.3 How Payment Calibration Affects the Measurement of Outcomes

In experiments involving real tasks and real effort, the measured outcomes are the total number of tasks completed (e.g., number of envelopes stuffed or puzzles solved correctly). Subjects must exert mental and/or physical effort to produce each unit of output measured in the experiment. An effective increase in incentives is one that is greater than the cost of effort and increases the value to the subject. Subjects offered one dollar per puzzle theoretically should solve a greater number of puzzles, on average, than subjects offered ten cents per puzzle.

Calibration of payments matters because tasks present different degrees of difficulty and subjects' aptitudes for completing different tasks vary. A subject who can easily complete a difficult task will likely respond to a relatively small incentive while a subject who finds the task daunting and must exert considerable effort will likely need a relatively large incentive to work harder and faster. For example,

most subjects will find stuffing envelopes is less physically demanding than delivering messages by bicycle. Consider two experiments: one offers $0.10 per stuffed envelope and the other offers $0.10 per message delivered via bicycle. Raising the incentive for stuffing envelopes from $0.10 to $1.00 per envelope will likely result in a greater increase in output than the same increase in incentive for bicycle messengers. Additionally, a subject who is a highly capable cyclist will likely be more responsive to the increase in incentive than a less-capable cyclist.

Incentives that are small relative to the difficulty of the task and ability levels of subjects will have little or no impact on outcomes measured in the experiment. Lacking a significant financial motive, subjects will determine the level of effort to invest using other criteria that the experimenter cannot control, introducing greater noise when measuring the treatment effects and diminishing the power of the experiment to detect treatment effects for a given sample size. Incentives that are unnecessarily large relative to the difficulty of the task and ability levels of subjects will not necessarily increase the power of the design to detect a treatment effect. Paying $20 per correctly solved puzzle, for example, might not motivate subjects more than $15, especially if they are already exerting the maximum effort they can to solve the puzzles. Efficient calibration of an incentive payment balances the marginal benefit to the experiment's power against the marginal cost of increasing the amount of payments made in the experiment.

Type of task may interact with payment in yet another way. The literature has paid considerable attention to the "crowding out effect" in real effort tasks (McGraw and McCullers 1979; Frey and Oberholzer-Gee 1997; Irlenbusch and Sliwka 2005; Pokorny 2008; Mellström and Johannesson 2008). Extrinsic (e.g. monetary) motivators can either further motivate or undermine a subject's intrinsic motivators. Intrinsic motivators can include a subject's believed civic duty or altruistic tendencies. Studies have found evidence that in some cases, extrinsic motivators such as incentive payments can undermine intrinsic motivators. Take for example, a stuffing envelopes task. For a fixed payment, a subject will have a baseline level of envelopes that they will stuff. When the price is increased to a fixed payment plus $0.1 per envelope, the subject may find the extrinsic incentive suitable to increase their effort in stuffing envelopes. On the other hand, if the task is giving blood, the subject may be willing to do it for altruistic reasons and lower the level of donation when payment is offered (Mellström and Johannesson 2008). Once that payment is involved and it becomes a transaction, a high level of payment may be needed to incentive the subject to sell blood. Subjects can experience the crowding out effect when small fees are paid for tasks they may have performed voluntarily, or when the wrong amount of incentive is paid (Uri Gneezy and Rustichini 2000). Furthermore, when incorrect monetary incentives are given, it can lead to "dysfunctional behavioral responses" in subjects (Prendergast 1999). These examples clarify the need for properly calibrating payments in every study, leading Gneezy and Rustichini (2000) to title their study "Pay enough or don't pay at all".

4.4 Review of Effect of Performance Incentives in Real-task and Real-effort Experiments

We identified three key studies that reviewed earlier economic experiments that had been designed to evaluate the performance of real effort task incentives. In the first, Bonner et al. (2000) reviewed 131 papers published between 1952 and 1998 in the accounting, psychology, and management literatures

that met their criteria with the specific goal of understanding the effect of performance incentives on performance of real effort tasks.[6] They concluded that quota schemes produced the best outcomes relative to the studies' control treatments, followed by piece-rates, tournaments (in which subjects competed to receive payment), and fixed rates. They argued that the likelihood of a financial incentive increasing performance declined with the complexity of the task.

In the second review, Camerer and Hogarth (1999) informally reviewed 74 studies that described substantially different levels of incentives used in detail. Some of those studies involved real tasks. The review was informal in the sense that it included studies that the authors had either previously read or that had been published in a select number of economic journals. For real effort tasks specifically, the authors concluded that incentives could improve performance for memory, simple problem-solving, and piece-rate clerical tasks. The results showed that the incentives either reduced performance or had no clear effect on other types of tasks. A third review, and meta-analysis, by Jenkins Jr et al. (1998) utilizing reported data from studies published in the 1960s, found that financial incentives increased quantity of output, but not quality across tasks.

We reviewed real effort tasks from the reviews mentioned, as well as collected additional real effort task experiments in the economics, psychology, and related literatures. We provide details about the types and scales of payments used, the results, and any issues associated with the method of analysis (such as confounding and deception problems) from those studies in Table 2. A comprehensive list of studies that were examined but could not be used for our analysis, typically due to a lack of available data on means and standard deviation of their outcomes, can be found in the appendix A.1.

Table 2: Studies Included in Analysis of Real Effort Tasks

Type/author	Subs	Txs	Subs/Tx	Incentive Type	Task	Payment	Incentive Effect
Allan et al. (2017)	40	2	20	F/P	Multiply	F: $6.34, P: $0.25/correct answer	=
Al-Ubaydli et al. (2015)	78	2	47,31	F/P	Stuff	F: $9 hr, P: $8 hr+$0.2/envelope	=/+
Bailey et al. (1998)	72	3	24	F/P, F/Q	Assemble	F: $20, P: $1.80/unit, Q: flat base of $17.50, plus bonus of $3 or $6	+
Bellemare (2010)	82	2	40-42	F/P	Data entering	$10 show up fee, F: $10, P: $0.10/entry	
Bracha et al. (2015)	88	2	44	P/P	Solve	P: $0.40/$0.80 correct answer	+
Cadsby (2007)	115	2	57-58	F/P	Find	F: $1.57/round, P: $0.14/word	+
Carpenter and Gong (2016)	207	6	34-35	F/P, P/P	Stuffing	F: $20/15 minutes, P: F + $0.5 letter/ F + $1 letter	+
Carpenter et al. (2010)	106	2	53	P/T	Stuff	P: $1 quality envelope, T: P + $25 winner	+

[6] The criteria for inclusion in the review were that the paper had been published in an English-refereed journal, used laboratory experiments or highly controlled field experiments, used adult human subjects, and addressed the effect of financial incentives on individual performance on a single task that was at least partially cognitive.

Type/author	Subs	Txs	Subs/Tx	Incentive Type	Task	Payment	Incentive Effect
Cason et al. (2010)	138	2	69	P/T	Add	P: $0.40/answer, T: $20 winner	N/A
DellaVigna and Pope (2017)	2226	4	540-556	P/P	Press buttons	P: $0/$0.01/$0.04/$0.1	+
Dillard and Fisher (1990)	27	2	13-14	F/Q	Decode	F: $3, Q: $2 + $0.54*(output over standard)	+
Dohmen and Falk (2011)	360	3	120	F/P/T	Multiply	F: $52, P:$0.22/correct answer or $0.22 shared by team, T: $0 vs $29	+
Erez et al. (1990)	16	2	8	N/Q	Type	Q: $0.25 for 0.1 better standard score (max $1.5)	=/+
Eriksson et al. (2008)	208	6	28-48	P/T	Add	P: $0.15/answer , T: $25 winner	N/A
Erkal, et al. (2018), exp 1	156	3	52	F/T, T/T	Encrypt	$10 show-up, F: 15 AUD, T: 25/60 AUD winner, 15 loser	=/+
Erkal, et al. (2018), exp 2	161	3	54	F/T, T/T	Encrypt	$10 show-up, F: 15 AUD, T: 25/60 AUD winner, 15 loser	+
Fatseas and Hirst (1992)	180	12	15	F/P	Decode	F: $5, P: $0.28 line (approx $5), Q: $1 participation, bonus for success	=/+
Fehrenbacher and Pedell (2012)	165	6	21-38	F/P/Q	Solve	F: $12.8, P: $0.3/anagram, Q: $35.8 target, $5.1 otherwise	N/A
Freeman and Gelber (2010)	468	6	78	F/P/T/T	Solve	Show-up fee: $13, F: $5 to all, P: $0.20/maze, T: $30 only winner, $15 /$7/$5/$2/$1	=/+
Friedl et al. (2018)	114	2	57	F/P	Slider task	F: 3.75 pounds, P: 0.03 pounds per slider	+
Frisch and Dickinson (1990)	75	5	15	F/Q	Assemble	Q: $4 + 0/10/30/60/100% above base pay for success.	+
Gächter et al. (2016), exp 1	20	2	10	P/P	Click	$4.6 show up fee, P: $0.01/$0.03	+
Gneezy and Rustichini (2000), exp 1	160	4	40	N/P, P/P	Answer	P: $0.03/$0.3/$0.86 correct answer	=/+
Goerg et al. (2017)	48	2	24	P/P	Slider task	P: $0.022/0.11 per screen	=
Greiner et al. (2011)	30	2	15	P/P	Data entering	P: $0.1, $0.25, $0.4	+

Type/author	Subs	Txs	Subs/Tx	Incentive Type	Task	Payment	Incentive Effect
Hamner and Foster (1975)	98	6	15-20	N/F/P	Transfer	F: $0.75, P: $0.05 scored survey	=/+
Libby and Lipe (1992)	117	3	38-40	F/P, F/T	Recall	F: $2 participation, P: F + $0.1/correct, T: F + $0.1/correct + top 5 $5.00 bonus	+
Pinder (1976)	80	4	20	F/P	Construct, assemble	F: $2.75, P: $0.05 piece	=/+
Pokorny (2008)	132	4	33	N/P	Count	$5 show-up fee, P: E0.01/E0.05/E0.5 point scored	+/-
Pritchard et al. (1977)	28	4	7	N/T	Solve	T: Best performer gets $5	=
Rubin (2016), exp 1	197	3	42-100	F/P, P/P	Add	P: $0.05, $0.25, $1, $3 per correct	+/-
Takahashi et al. (2016)	145	5	24-34	N, F/F, P/P	Click, solve	N: no pay, P: $0.005, $0.02 for circles, $1/$4 for puzzles.	=/+, +
Terborg and Miller (1978)	60	6	10	F/P	Assemble	F: $2.50, P: $0.4 model	+
Tonin and Vlassopoulos (2013)	104	2	52	P/P	Data entering	P: $0.03, $0.03/$0.06/$0.10/$0.13	+
Van Dijk et al. (2001)	79	3	24-28	P/T	Find	P: variable, T: $0.6 win, $0.15 lose, $0.37 tie	N/A
Vandegrift et al. (2007)	180	4	45	P/T	Forecast	P: variable, T: $4.5 winner, $2.25/$1.5/$0.75/$0	N/A
Yukl et al. (1972)	15	3	5	V/P	Grade	F: $1.50/h, P: F + 25c/sheet, Variable-Rate: F + 50-50 chance 25c/sheet, Variable-Rate: F + 50-50 chance $0.5/sheet	=/+

Notes: No monetary rewards (N), Fixed (F), Piece-rates (P), Tournaments (T), and Variable-rate (V) payments. Incentive type columns shows the study's compared incentive types.

Figure 1 and Figure 2 present coefficients of variation calculated for the studies that met our criteria for inclusion (no confounding effects). The analysis shows that the coefficients of variation decrease as the level of incentive increases in the piece-rate, quota, and tournament experiments.

Observationally, the piece-rate figure presents a pattern in which the coefficient of variation decreases as the payment for performance increases. The relationship between the coefficient of variation and

payment may not be linear, but the volatility in outcomes appears to shrink as payments increase. A caveat to the results is that the studies are not homogenous in terms of design. For example, figure 4.1 shows an outlier at $1 with a coefficient of variation of 0.87. The particular experiment set out to test the importance of a mission on a subject's effort, and the effect of monetary incentives given aligned or misaligned intrinsic incentives. The specific data point showed participants who were asked to work for candidates of the opposite spectrum of beliefs, thus having very low intrinsic motivation for the task. An additional point to note is that data becomes sparse after a payment of $0.5 per output so that it is not possible to know whether the data points obtained from the available studies are truly representative of the effect of such payments on performance.

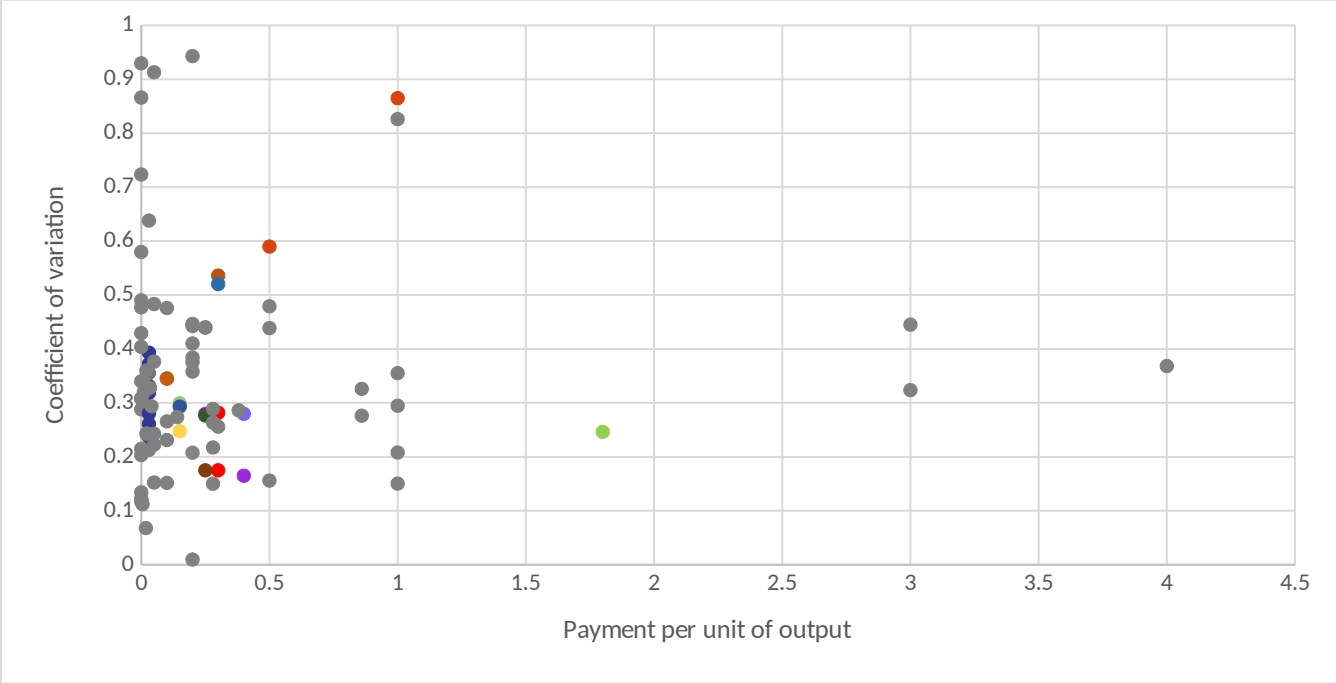


Figure 1: Coefficients of Variation of Monetary Incentives for Experiments Involving Piece-Rate Payments

Sources: Libby and Lipe 1992; Fatseas and Hirst 1992; Bailey, Brown, and Cocco 1998; Yukl, Wexley, and Seymore 1972; Hamner and Foster 1975; Pinder 1976; Terborg and Miller 1978; Rubin, Samek, and Sheremeta 2016; Gneezy and Rustichini 2000; Al-Ubaydli et al. 2015; Pokorny 2008; Goerg, Kube, and Radbruch 2017; Takahashi, Shen, and Ogawa 2016; DellaVigna and Pope 2017; Dohmen and Falk 2011; Allan, Bender, and Theodossiou 2017; Cadsby, Song, and Tapon 2007; Freeman and Gelber 2010; Fehrenbacher and Pedell 2012; Tonin and Vlassopoulos 2013; Carpenter and Gong 2016; Bellemare and Kröger 2007; Greiner, Ockenfels, and Werner 2011; Gächter, Huang, and Sefton 2016; Eriksson, Poulsen, and Villeval 2008; Cason, Masters, and Sheremeta 2010; Bracha, Gneezy, and Loewenstein 2015; Carpenter, Verhoogen, and Burks 2005; Friedl, Neyse, and Schmidt 2018.

Similarly, observationally the figure on quota and tournament payments appears to show a decline in the coefficient of variation as the winner's bonus or quota increases. The data for these types of payment schemes is even sparser than that for piece-rates, but the range in the coefficient of variation appears to shrink with higher payment values.

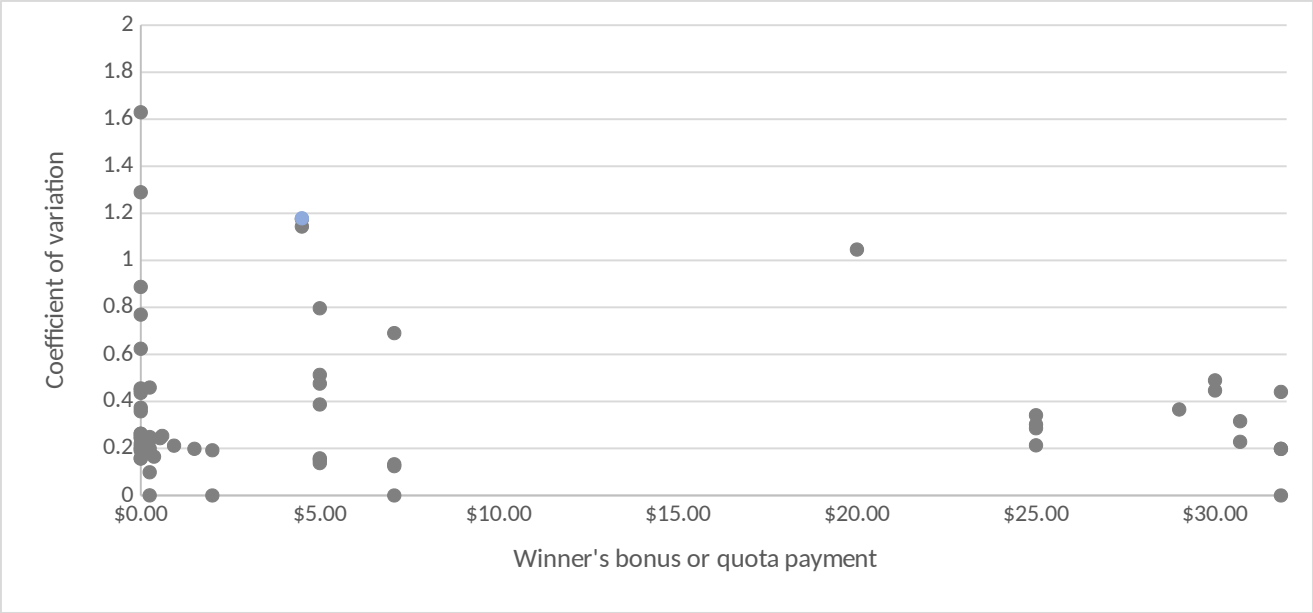


Figure 2: Coefficients of Variation of Performance When Monetary Incentives Were Provided for Experiments Involving Quota and Tournament Payments

Sources: Erkal, Gangadharan, and Koh 2018; Libby and Lipe 1992; Carpenter, Verhoogen, and Burks 2005; Dillard and Fisher 1990; Erez, Gopher, and Arzi 1990; Dohmen and Falk 2011; Freeman and Gelber 2010; Van Dijk, Sonnemans, and Van Winden 2001; Cason, Masters, and Sheremeta 2010; Eriksson, Poulsen, and Villeval 2008; Fehrenbacher and Pedell 2012; Irlenbusch and Ruchala 2008; Vandegrift, Yavas, and Brown 2007; Pritchard, Campbell, and Campbell 1977

Table 3 shows the count of studies for select comparisons which show either positive, ambiguous, negative or no effect on performance. There are several reasons why a study may be characterized by an ambiguous effect. For example, Gneezy and Rustichini (2000) designs piece-rate payment schemes of about $0.03, $0.3, and $0.86 per correct IQ question answered. Increasing payments from $0.03 to $0.3 for each correct answer increases performance, but increases from $0.3 to $0.86 does not. Also of interest is the negative effects shown in the figure. None of the studies that have sufficient information for analysis showed an unambiguous negative effect from increased payments. Three studies show ambiguous effects that include a negative effect from increasing payments. For example, Rubin et al. (2016) find that higher incentives increase quality, but decrease quantity, demonstrating that the output being measured is of great importance.

Table 3: Effect of Performance on Output

	Positive	Ambiguous (=/+)	Ambiguous (=/-)	Ambiguous (+/-)	No effect	Negative
Memory	1	0	0	0	0	0
Production and Clerical	5	4	0	0	0	0
Problem Solving	10	4	2	1	2	0

In general, economic studies have mostly found positive correlations between larger financial incentives and performance of tasks. Araujo et al. (2016), for example, found a small but positive correlation between the number of tasks completed correctly (properly positioning a slider at the midpoint) in a between-subject design that compared piece-rates of 0.5 cents, 2.0 cents, and 8.0 cents. Dalton, Gonzalez Jimenez, and Noussair (2016) compared low and high piece-rates and found that the subjects' output under the high piece-rate was 7% higher than their output under the low piece-rate.

4.5 Limitations of Excluded Studies

Our analysis excluded many studies that were included in previous surveys of the literature, but included all studies that had available data for analysis.

Particularly, many of the studies included in Bonner et al. (2000) did not report relevant statistics such as the standard variation and did not cluster the results per session.[7] Table A.1 in the appendix details the reason for the exclusion of studies reviewed.

Confounding of results is another major problem when studies mix financial incentives with other kinds of motivation (e.g., comparing financial schemes while also directing subjects not to expend much effort). Some of the studies we reviewed also compared results with no financial incentive to results from applying negative (punishment) and positive (reward) financial incentives.

Aside from punishment, some early experiments also used unethical treatment of subjects and deception. Bevan and Turner (1965), for example, analyzed the effect of variable piece-rate payments versus no monetary reward on performance of a vigilance task, but the incentive payment was mere pennies per piece while errors were punished by delivery of mild electric shocks. Glucksberg (1962) deceived "control" group members by telling them that they were participating in a pilot experiment to prepare for future experiments. And in an odd experiment design, Pollack and Knaff (1958) punished subjects for failing tasks by subjecting them to a 0.5-second blast of a GMC truck horn positioned 18 inches in front of the subject.

[7] It is standard practice in experimental economics to cluster the errors (typically at the session level); otherwise, the errors are downward-biased.

5.0 Effects of Performance Payments by Experiment Methodology

In this section, we review the effects of performance payments on four types of experimental methodologies commonly used in ERS experiments: auctions, multi-player games, market experiments, and risk preference elicitations. These methods differ from real task and real effort experiments in two ways. First, these methods use stylized settings that parallel a real-world situation. The extent of the stylization may vary by method and study, but experimentalists often design the method to minimize the likelihood of subjects having intrinsic motivations to guide their decisions (see discussion of framing effects in Section 6.4). Second, these methods typically do not impose real costs on subjects. They rely on costs induced within the experimental setting. Minimizing the likelihood of subjects making decisions according to intrinsic motivations and inducing costs within the experimental setting both help reduce the likelihood of confounding factors interacting with the experimental treatment in uncontrolled ways.

Calibration of incentive payments is likely to be more important for methodologies that rely on stylized tasks and induced costs than for real task and real effort experiments. With real task and real effort experiments, the setting is more familiar and subjects may be able to rely on experience with the task from outside of the experiment to formulate strategies and make decisions. For methodologies using stylized tasks and induced costs, subjects may need to exert extra cognitive effort to determine their optimal decisions within this unfamiliar setting.

For each methodology, we explain why the method is relevant to inform USDA policy-making, explain how the method works, describe how payment calibration levels impact the outcome measured, review experimental norms for setting payments, and present the results from our analysis of the impact of performance payments on outcomes.

5.1 Auctions

5.1.1. Why this method is relevant for informing USDA policy-making

Willingness to pay (WTP) is the amount of money an individual is willing to spend in order to purchase a good. Similarly, willingness to accept (WTA) is the amount of money an individual requires in order to sell a good. WTP and WTA can be used as alternative methods of valuing a good (instead of relying on a market price), or to infer a value for a good when no market prices are available. Many experimental, quasi-experimental, and survey methods can be used to estimate individuals' willingness to pay (WTP). In this section, we focus on experimental auctions for eliciting WTP or WTA.

Auctions studies have been used to inform a wide-variety of USDA policymaking, including fire hazard reduction programs (69 FR 35576), prevention of foodborne diseases (67 FR 15169), outdoor access for chickens (7 CFR 205), and agricultural quarantine and inspection services (7 CFR 354). For example, auction experiments have been used to identify WTP for beef with country-of-origin labeling (see Loureiro and Umberger, 2007; Lim et al., 2013) in order to estimate demand curves for a premium-priced inspection-differentiated product. Auctions have also been used to investigate behavior under novel policy designs, such as contracts for environmentally beneficial land uses and production practices offered by a conservation program and allocated through some form of auction (see Fooks et al., 2016; Wallander et al., 2017; Messer et al., 2017). In these studies, the auctions were used to discover what

prices farmers would need to participate in conservation programs or what combination of environmental benefits could be purchased for a specific payment amount.

Auction-like markets for conservation programs are of particular interest to USDA when a budget or statutory limit to the number of contracts awarded limits the number of participants who can be selected. Auction formats potentially allow agencies to enroll a greater number of participants and provide a greater social benefit by reducing incentives to overbid (Ferraro, 2008; Reeson and Whitten, 2014). They can also contribute to identifying the optimal allocation of the contracts (Latacz-Lohmann and van der Hamsvoort, 1997; Schilizzi and Latacz-Lohmann, 2007).

Other Methods of Estimating Willingness-to-Pay

In addition to auctions, ERS uses revealed-preference and stated-preference choice experiments to estimate willingness-to-pay for goods and services. Revealed-preference experiments ask subjects to make choices in actual markets, and the resulting data can be observational (with or without the experimenters' influence) or self-reported. These experiments typically are used to estimate the value individuals ascribe to items when they reveal their preferences for those items by their choices. The data collected from non-hypothetical experiments (e.g., conducted in the field) are often used as a proxy for observed data in natural settings (Fifer et al., 2014). Stated-preference experiments typically are used to elicit information in hypothetical markets and the techniques, such as contingent valuation and discrete-choice experiments, are applied to estimate the monetary value of commodities that lack a market (Hanemann, 1994; Hensher, 1997; Ryan and Gerard, 2003; Ryan and Gerard, 2008; Smith, 2003).

We omit choice experiments from our main analysis because the payments to participants in those types of experiments usually are not contingent on choices made in the game. Instead, participants usually receive a flat fee as compensation for the time they provide (ranging from the hourly minimum wage to about $50). However, a major criticism of using stated-preference methods to study choices and elicit individual WTP for goods is that the choices are made in hypothetical markets. Critics have questioned whether the responses accurately represent "real world" behavior (Diamond and Hausman, 1994) or suffer from hypothetical bias (Fifer et al., 2014). These concerns about hypothetical bias in choice experiments parallel concerns about hypothetical bias in auctions.

5.1.2. How auctions work

Experimental auctions are structured similarly to auctions in the real world. Subjects can take the role of bidders and/or sellers. The outcomes measured can include the market clearing price; the number of trades executed at the market clearing price; the distribution of bids/asks posted by subjects; and/or consumer, producer, or social surplus. Depending on the research question, the experimenter may choose to control different aspects of the auction, such as the number of bidders and sellers, the mechanism for setting the winning bid, the number of units auctioned, and/or the valuations of bidders/sellers. Recent surveys of this broad literature are given by Kagel and Levin (2011) and Dechenaux *et al.* (2015). A more narrowly targeted survey that focuses just on conservation auctions is given by Schilizzi (2017).

Experimental auctions are classified based on (1) whether subjects are bidders (a valuation auction), sellers (a reverse auction), or both (a double auction); on (2) whether the value of the auctioned good matters only to individual bidders (private value auction) or if there is some common value across bidders (public goods auction); or on (3) the mechanism used to set the winning bid. For a valuation auction, the mechanism determines which buyers will receive the item purchased and how much those buyers need to pay. For a reverse auction, the mechanism specifies which sellers will have completed sales and what prices they will receive. Some commonly used mechanisms include:

- First price sealed bid auction: In this format, each participant places a single private bid. The highest bidder wins the auction, receives the item, and pays the highest price bid.
- Vickrey (Vickrey, 1961): Vickrey auctions are also known as sealed-bid second-price auctions. In this format, each participant places a single private bid. The highest bidder wins the auction, receives the item, and pays the amount bid by the next highest bidder.
- Random nth price (Shogren et al., 2001): This auction is similar to a Vickrey auction except that a randomly chosen "nth" bidder sets the market price (and does not purchase the item). All of the bidders who offered more than the nth bidder win the auction, purchase the item, and pay the nth bidder's price. The nth bid can either be pre-announced or determine randomly after all of the bids have been submitted.
- English auction: This is an "open cry" ascending price auction where bids are made publically. In this format, the auctioneer announces a starting price, and then bidders place a series of ascending bids until no additional bids are made. The highest bidder wins the auction, receives the item, and pays the highest price bid.
- Dutch auction: This is also an "open cry" descending price auction where bids are made publically. In this format, the auctioneer announces a starting price and then starts lowering the price in set increments until at least one bidder makes a bid. The bidder wins the auction, receives the item, and pays the amount bid.

The type of mechanism used affects the participants' incentives to bid their true valuations. A well-designed auction is incentive-compatible – meaning that participants' best strategy to win the auction is to bid their true "value" for the product instead of strategically over-bidding or under-bidding. Incentive-compatible mechanisms also ensure that winning participants never pay more than their maximum WTP and winning sellers never earn less than their minimum WTA. Experiments that compare the performance of different auction mechanisms against each other or theoretical benchmarks generally induce values for participants' WTP/WTA in order to have more control over bidding behaviors and improve the precision of the estimated treatment effect.

Not all mechanisms are incentive-compatible. Notably, the most commonly used first-price sealed-bid auctions, which solicit private bids and award the contract to the lowest bidder, are not incentive-compatible. Firms have an incentive to under-bid relative to their true valuations to obtain the contract. If a competing firm's valuation is v_i and there are N firms, each firm will bid $b_i = ((N - 1) / (N))v_i$; bidders have the incentive to "shade" their bids to attempt to extract a profit of $(1 / N)v_i$. When the number of bidders is small, the profit extracted can be large (up to one-half of the item's value). As the number of bidders increases, the bids approach the firms' true valuations.

Many auction formats in which the bidders set their own prices are also not incentive-compatible. The classic English auction in which bids are publicly announced until a single bidder remains is theoretically incentive-compatible but a number of studies have suggested that dynamic public bidding tends to lead to over-bidding. This has been variously attributed to competitiveness and a desire to "win" (Heyman et al., 2004), inattentiveness (Malmendier and Lee, 2011), and to violations of the assumption underlying such auctions that participants' values are independent. If they are not independent, participants could take cues from other bidders, a form of the "winners curse" in which the winner ends up overpaying in common-value and correlated-value auctions (Thaler, 1988).

Conservations auctions fall into the category of non-incentive compatible, sealed-bid auctions, which are a common form of government auctions. Examples of such auctions include forward auctions for commodities such as radio frequency spectrum, timber on federal lands, and fisheries quotas or reverse auctions for things such as water rights, emissions credits, and conservation program contracts. Conservation programs involve farmers or other land owners competing to enroll in a program that will provide annual payments in exchange for establishing and maintaining conservation cover on environmentally sensitive land. When interest in these programs is greater than available funds, program agencies can either use a priority system (first-come-first-served) or an auction (ranking) to choose which offers to accept and which to reject. When framed as a way for governments to acquire the private provision of environmental services, conservation auctions are similar to other government procurement auctions. However, conservation auctions have three features that, in combination, make them a unique setting. First, they are multi-unit auctions that can involve thousands of individuals winning a single round of the auction. Second, they are quality differentiated, since there is a great deal of variation in the environmental benefits of enrolling different parcels of land. Third, there may be agglomeration effects in the benefits such that the program managers would like to enroll contiguous clusters or corridors of land if at all possible.

The first two features - the fact that most conservation programs must accept multiple units that differ in both quality and price – imply that conservation auctions generate rents for accepted offers. Typically these are referred to as information rents because they arise from the combination of information about costs (which is asymmetric, known only to the sellers) and about quality (which is typically symmetric) (Ferraro 2008). For this reason, most experiments on conservation auctions test alternative auction designs for differences in the levels of rent for given amounts of conservation. The idea is that program agencies can implement more cost effective programs if they use auction designs that minimize rent (i.e.: price discriminate). Some studies incorporate the third feature of conservation auctions by incorporating spatial targeting, which creates an explicit tradeoff between higher benefits and lower rents.

5.1.3. How payment calibration affects the measurement of outcomes

The outcomes measured in an auction can be the value of the winning bid(s), market clearing price, quantity traded, and/or buyer or seller surplus. Payments for auction experiments are designed to incentivize participants to exert effort to determine their optimal bids. In valuation auctions, participants are endowed with a quantity of money to use for bidding in the auction. The value of the endowment not bid is typically the subject's money to keep. In conservation auctions and other auctions used for purpose of testing theoretical predictions (i.e. auction mechanism design), subjects' earnings are the

based on the surplus they earn from buying or selling goods won in the auction. Higher incentive payments mean greater surplus for each transaction, providing incentives for participants to work hard to win the auction and disincentives to bid at random.

Endowment effects are an important issue when designing the experiments. The concept of an endowment effect arose from discrepancies observed between choices in experiments framed WTA versus WTP. Numerous studies have shown that endowing individuals with a payment to use to bid in an auction changes the individual's behavior relative to the participant having to pay "out of pocket" (Ackert et al, 2006; Jacquemet et al, 2009; Corgnet et al, 2014). Though there can be a sense of reciprocity with the administrator, the endowment effect is most often attributed to participants accounting for the payment mentally as a windfall rather than as normal income. Efforts used to avoid endowment effects have included prepayments in which the participants receive the payments when they sign up for the experiment several weeks prior to the session, and real-effort tasks have been used to make participants "earn" the money at the beginning of the session so they treat it more like earned income.

A second issue is use of induced versus "homegrown" values. An induced value is one that is assigned by the experimenter to each participant. It is a useful methodology because it gives the experimenter greater control over the distribution of values and allows observation of the actual underlying value so the results can be easily compared. Homegrown values are produced through competition between participants for an item for which they have developed an intrinsic value, such as a coffee mug or a food item. Some economists have expressed concern that individuals' behaviors in response to induced values differs from situations that require them to formulate and express a value themselves (Cummings et al., 1995; Murphy et al., 2010). In recent years, induced values have been used almost exclusively in such experiments (e.g., Schilizzi and Latacz-Lohmann, 2007)

5.1.4 Norms for incentive payments for auction experiments

Conservation auctions tend to be complex for bidders so incentives for their participation must adequately compensate them for the time and cognitive effort required (Harrison, 1989; Merlo and Schotter, 1992). Among the conservation auctions reviewed for this analysis, average payments to participants, almost always undergraduate students, ranged from about $15 to $38. We normalized these based on the time involved in each study and found that the average hourly rates ranged from $11 to $25 per hour, not adjusted for inflation. Because conservation auctions are complex mechanisms, it is not possible to establish an expected return from bidding at random for all studies reviewed. Therefore, we analyze these studies using the average hourly rate as opposed to the expected return to exerting effort.

In experiments involving valuation auctions, participants typically receive payment in cash and/or goods. They generally receive a cash show-up fee and can use some or all of that money to buy items won in the auction. Participants are rarely allowed to bid more than the amount of the show-up fee so they cannot end up owing money to the experiment administrators (though there are some exceptions, such as in Kecinski et al. (2017)).

When an auction involves a relatively expensive item, the payment can be scaled down by offering it in a lottery instead of a direct purchase. For instance, Fooks et al. (2017) sold tickets for a 1-in-10 chance at a night's stay at a beach resort for a maximum bid of $30 using a BDM mechanism. A lottery allows one to

collect data on higher-cost items, but additional assumptions must be made regarding risk preferences to interpret the results.

Valuation auction experiments do not necessarily require show-up payments; they only require that participants pay for any items they win. Ideally, the size of the show-up payment does not matter since we are trying to measure values that should not be affected by relatively small changes in wealth. Indeed, Rutstrom (1998) compared the results of auctions with and without initial payments and variations in the amount of payment and found no evidence that the presence or amount of a payment significantly affected the bids.

5.1.5 Results for the impact of performance payments on conservation auctions

In looking for experimental studies of conservation auctions, we identified thirty candidate studies. More than half of these studies did not include sufficient information on payments to participants or estimates of relevant treatment effects to be included here. Eleven studies provide sufficient information and involved estimation of a treatment effect on either average rent or total cost. Some of these studies find that withholding information on ranking can reduce information rents (Cason and Gangadharan 2004, Banerjee et al. 2015), while other studies find that withholding ranking information can also reduce benefits (Conte and Griffin 2007). Several studies suggest that pay-as-bid (discriminatory pricing) can reduce costs (Cason and Gangadharan) relative to uniform pricing, but that ordering can reverse if contract compliance decisions are taken into account (Kawasaki et al. 2012). Other key issues covered in these studies include the way in which the dynamic of repeated auctions can improve net benefits even while increasing rents (Fooks et al. 2015), the prevalence of adverse selection in these auctions (Arnold et al. 2013), incentives for offer quality improvement (Banerjee et al. 2018), the impact of excessively restrictive bid caps (Hellerstein et al 2015), the impact of using benefit-cost ratio ranking (Iftekar and Tisdell 2014, Fooks et al. 2015), multiple, interacting auctions (Tisdell and Iftekar 2013), and the role of communication and trust between participants and program administrators (Vogt et al. 2013).

In order for these studies to detect differences in rent extraction under the alternative auction structures, participants must be attempting to maximize their rents. If participants are not trying to maximize rents and just want to "win" the auction, then they would simply offer their land to the program at their minimally acceptable price and no rents would be generate. If participants were simply randomly submitting bids and not trying to maximize rents, then there would be no difference in rents between treatments. To insure that participants are focused on the problem of extracting rents from the auction, most experimental studies provide payments related to the rent that they can extract. This approach gives participants a task in which they can increase the probability of having their offer accepted by lowering the price at which they are willing to sell and so the must balance the reduce rent against the increased probability of acceptance. The information that participants must take into account in each round of an auction includes: 1) Their opportunity cost of enrolling, which is the return they get if their offer is not accepted, 2) the rank of their "land" relative to the other participants, and 3) the likely competitiveness of the auction in terms of the expected acceptance rate. In this setting, participants that get an "endowment" of lower costs or higher ranking are expected to extract more rents.

Since most studies used a two-by-two treatment design, there are usually three treatment dummies (and one control, the constant) or two if they aren't testing the interaction. For purposes of the chart below, we pick the treatment effect with the median statistical significance within each study.

Our primary finding is that the largest coefficient of variations on the treatment effect estimates occur for low and moderate average hourly payment rates (Figure 3). This pattern is consistent with the idea that higher payment rates could lead to increased attention to the task of extracting rents from the auction and improved statistical power for detecting differences in rent extraction across auction designs. However, some of the lowest coefficient of variations occur on studies with some of the lowest average hourly payment rates. A number of factors could influence this pattern.

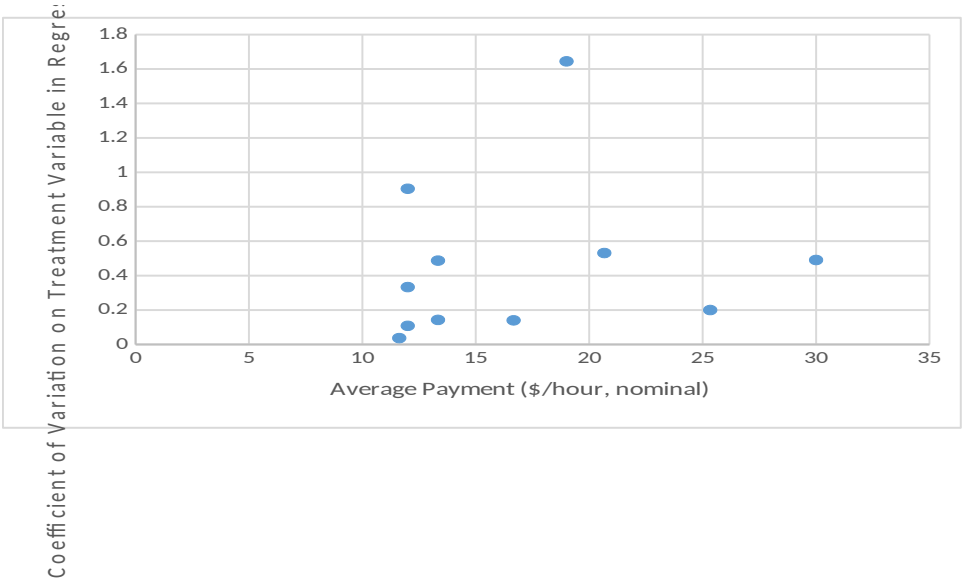


Figure 3: Participant Payments and Models of Information in Experimental Conservation Auctions

In the studies identified in the above charts, participants were asked to make offers in multiple auctions. Total number of auctions ranged from a low of 8 to a high of 65. In about half of the studies, auctions included multiple rounds, which are opportunities to revise offers within an auction.[8] Usually a participant is given a single item (e.g.: a field) on which to make an offer, but in some cases participants were given multiple items. The combination of auctions, rounds, and multiple items mean that over the course of a single session a participant could be making a lot of offer decisions. The most involved experiments involved 91 (Banerjee et al, 2015), 108 (Cason and Gangadharan, 2004), and 130 decisions (Fooks et al., 2016).

[8] The terminology used in this literature is extremely inconsistent. The studies use the terms sessions, rounds, periods, trials, and eras. Typically the term "session" refers to a single experimental session, or group of participants doing one full run of the experiment. Here we use the term "round" to refer to one full auction, but several of the papers refer to these as "periods" and use the term "rounds" to refer to the revision opportunities within an auction.

There are two caveats with this analysis. First, there is a lot of variation in the auction structures that are tested. This will of course impact the size of the treatment effect and therefore the size of the CV. So a critical assumption of our research is that the variation in the CV due to payment levels is orthogonal to the variation due to the hypothesized treatment effects. Second, there is a lot of variation in how the studies calculate their standard errors. The experimental designs vary in the number of auctions, the number or revision periods or rounds, and whether treatment are within or between subjects. The studies differ in how they manage all of this variation in terms of clustering standard errors and testing models that include various interactions. Again, we need to assume that the variation from these econometric decisions are orthogonal to the variation from the payment levels.

We made one robustness check on the chart above. We normalized the CV on sample size to account for the idea that part of the variation in CV is coming from statistical power. We used the number of participants at the "N," but as with the standard errors issue above, the studies vary a lot in what they actually use as the "N" in their regressions. Since the studies tend to have similar sample sizes, this variation did not change the chart significantly.

5.2 Multi-player Games

5.2.1. Why this method is relevant for informing USDA policy-making

Experimental economists use multi-player games to measure cooperative behaviors, trust, and the extent to which outcomes affecting other people matter for an individual's decision-making (i.e., other regarding preferences). Cooperation, trust, and other regarding preferences impact many aspects of agricultural enterprises and markets, including farm ownership and governance (Steier, 2001), participation in agricultural co-ops (Chaddad and Cook, 2004; Osterberg and Nilsson, 2009), organization of agricultural markets (Boehlje, 1999), and marketing and labeling of agricultural products (Toler et al., 2009; Briggeman et al., 2010). Attitudes about cooperation, trust, and other regarding preferences influence decisions about resource conservation (Adams et al., 2003; Velez et al., 2009), environmental management (Venkatachalam, 2008; Ostrom, 2010), and agricultural policy (Ellison et al., 2010). Public good games and threshold public good games (aka. provision point mechanisms) are highly relevant to the work of USDA as they are fundamentally related to voluntary participation in agri-environmental programs.

Multi-player games are used in the academic literature to test economic theories, such as whether individuals always behave in a rational and self-interested way. These games can be varied in numerous ways: whether players act simultaneously or sequentially, whether they have perfect information about the game and others' pay-offs, whether the game involves a fixed total pay-off (zero-sum) or pay-offs that are affected by players' actions, and whether there are (weakly) dominant strategies. Subjects often play the same basic game repeatedly over many rounds, with treatments applied between subjects that change aspects of the design, the stakes, the type of participants and cultures, rules of play, and/or the number of rounds. For this paper, we focus on the effects of the size of the players' stakes on the outcomes.

5.2.2. How multi-player games work

There are many types of multi-player games used by experimental economists to measure cooperative behaviors, trust, and other regarding preferences. In this report, we focus on the methods most relevant for USDA policy-making: the ultimatum game, the dictator game, the voluntary contribution mechanism, and trust games.

1. Ultimatum game (UG): In an ultimatum game, player 1 is given a sum of money and can offer some (a percentage) or all of it to player 2. If player 2 accepts the offer, both players retain the share of money determined by player 1. If player 2 rejects the offer, both players end up with nothing, thereby forfeiting their potential earnings. Following standard economic theory, player 1 can infer that player 2 will be better off with any fraction of the money than with none and thus be willing to accept even small offers from player 1. Therefore, in theory, player 1 will offer a very small fraction of the money to player 2.
2. Dictator game (DG): The dictator game (DG) is a version of the UG in which player 2 cannot reject the offer. Consequently, player 1 is a dictator and controls the amount of pay-off s/he receives in the game, allowing one to test whether player 1 (the dictator) will give any money to player 2 in various circumstances. Since player 1 has no fear that the share offered will be rejected, s/he can offer nothing and retain the entire sum of money.
3. Voluntary Contribution Mechanism (VCM): In a voluntary contribution mechanism game – a type of public good games (PGGs) - each of *n* players simultaneously decides how much of an endowment to contribute to a public good (a common pot), and the players retain the rest of their individual endowments. Given the sum of the players' contributions, the experimenter chooses a percent "return" on their investment (e.g., a 100% return that doubles the total investment or a smaller percentage such as 50% or 10%). The experimenter then sums the contributions and the return, divides that sum by the number of players, and distributes a share to each player regardless of the player's initial contribution. Thus, the players earn the share of the endowment they kept plus the share of the common pot. A cooperative strategy in which every player contributes 100% of the endowment produces the greatest earnings. However, the contributions are voluntary, and if other players contribute to the common pot, any single player's best individual strategy is to give nothing.
4. Trust games: In trust games (also known as centipede games), player 1 decides how much of an initial monetary endowment to give to player 2 knowing that the percentage player 2 receives will be increased by a factor set by the experiment administrator (e.g., doubled or tripled). Player 2 then decides how much of the money to return to player 1 (similar to a DG). By choosing to share with player 2, player 1 is trusting that player 2 will return at least the amount of money given. When player 1 does not trust that player 2 will return the amount offered, player 1's rational strategy is to keep all the money in step 1 and give none to player 2. However, experiments have demonstrated that player 1 often tends to trust player 2 so the "no trust" strategy does not always occur.

In each of these games, players are assumed to follow a strategy that will give them the greatest amount of pay-off (money) contingent on the strategies likely to be chosen by the other players. These decisions may not result in the highest possible payout for individual players. Rather, these decisions result in the best possible outcome in the sense that, all else being equal, each player is maximizing his/her own payoff and no player would choose to make a different decision (a Nash equilibrium).

5.2.3. How payment calibration affects measurement of outcomes

Error in a multi-player game occurs when players make decisions in a game that are inconsistent with their true preferences. The magnitude of the error scales with the size of the available endowment. Thus, higher payment levels (larger endowments) provide greater incentives for players to reveal their true attitudes.

For example, consider a player whose true preferences are to claim 80 percent of the endowment in a dictator game. This player would prefer to keep $2.40 from an endowment of $3.00 and $24.00 from an endowment of $30.00. If this player makes an error and keeps only 75% of the endowment, the magnitude of this error would be $0.15 when the endowment is $3.00 and $1.50 when the endowment is $30.00.

The opportunity for players to make errors also depends on the complexity of the endowment. It is easiest for players to calculate the share of the endowment they would like to retain for endowments that are round numbers and multiples of tens (e.g., an endowment of $10.00 as compared to an endowment of $11.62).

5.2.4 Norms for incentive payments in multi-player games

Incentive payments can range from several dollars to several hundred dollars, although the range of incentives used in developing countries tends to exceed the range of incentives used in developed countries.

5.2.5 Results for the impact of performance payments on outcomes

Due to the large literature on multi-player games, we focus our review on the studies that explicitly test for differences in performance under multiple incentive schemes. We include studies from developing countries as these studies test a greater range of incentives than the majority of studies from developed countries. Because only a small number of studies report enough information to calculate a CV for the study's outcomes, we report only summaries of the studies' conclusions. The results are shown in Table 4.

Fifteen of the games found no statistically-significant change in outcomes with higher stakes. Fourteen games found statistically-significant changes in outcomes consistent with theoretical predictions (i.e. lower offers and/or fewer offer rejections). One more game found changes in outcomes consistent with theory either for only one of the populations sampled. Another game found changes in outcomes consistent with theory when the game was played multiple times but not when the game was played once. Two games found statistically-significant changes in outcomes are counter to theoretical predictions.

Of the twelve games where we could assess the effect of higher stakes on the variance in outcomes, three studies found no impact on the variance of outcomes, five studies found less variance in outcomes with higher stakes, and three studies found increasing variance or CV with higher stakes. Additionally, one study found decreasing variance with higher stakes but only for real payments, not hypothetical

payments. All but one of the studies showing no effects of incentives on outcome variance or CV involved hypothetical payments.

Shifting from no monetary stakes to positive monetary stakes tended to have a larger impact on behavior than going from low stakes to high stakes, consistent with a hypothesis of diminishing marginal returns to incentives. Other design elements of the experimental games sometimes had much larger impacts on the players' choices than the stakes did. However, many of these studies used relatively small samples and could have been underpowered to observe the true effect of monetary stakes on outcomes for these games.

Table 4: Summary of Results for Multi-Player Game Experiments

Study	**Description**	**Number of subjects**	**Game Type**	**Comparison**	**Effect of Incentives**	**Effect on Outcome Variance**
Amir, Rand, and Kobi Gal (2012)	Mturk Participants	756	Ultimatum	Hypothetical vs $1	Larger offers	No change
Amir, Rand, and Kobi Gal (2012)	Mturk Participants	756	Dictator	Hypothetical vs $1	Lower offers	Less variance
Amir, Rand, and Kobi Gal (2012)	Mturk Participants	756	Trust	Hypothetical vs $1	No change	No change
Amir, Rand, and Kobi Gal (2012)	Mturk Participants	756	Public-Good	Hypothetical vs $1	No change	No change
Andersen et al. (2011)	Indian villagers	458	Ultimatum	$0.41 vs $4.1 vs $41 vs $410	Lower offers, less rejection	-
Cameron (1995)	Indonesian students	282	Ultimatum	$10-15 vs $80-120	No change	Decreased in real treatments but not in hypothetical treatment
Carpenter, Verhoogen, and Burks (2005)	US students	39	Ultimatum	$10 vs $100	Lower offers	More variance
Carpenter, Verhoogen,	US students	40	Dictator	$10 vs $100	No change	Less variance

Study	Description	Number of subjects	Game Type	Comparison	Effect of Incentives	Effect on Outcome Variance
and Burks (2005)						
Cherry, Frykblom, and Shogren (2002)	US students	174	Dictator	$10 vs $40	No change	-
Diekmann (2004)	Students	69	Dictator	$2.60 vs $6.50 vs $10.40	No change	-
Engel (2011)	Meta-analysis of 129 studies	20813	Dictator	$0 up to $130	Lower offers	-
Forsythe et al. (1994)	US students	206	Ultimatum	$0 vs $5 vs $10	Less rejection	-
Forsythe et al. (1994)	US students	206	Dictator	$0 vs $5 vs $10	Lower offers	-
Gillis and Hettler (2007)	US students	42	Ultimatum	$0 vs $10	Lower offers	Less variance
Gillis and Hettler (2007)	US students	80	Public-Good	Hypothetical vs real	No change	Less variance
Guth, Schmittberger, and Schwarze (1982)	German students	42	Ultimatum	1.8 DM vs 18 DM	No change	-
Henrich (2000)	US students	81	Ultimatum	$10 vs $160	No change	Less variance
Hoffman, McCabe, and Smith (1996)	US students	196	Ultimatum	$10 vs $100	No change	-
Holm and Nystedt (2008)	Swedish students	-	Trust	Hypothetical vs random lottery payment	Larger offers	-
Johansson-Stenman, Mahmud, and Martinsson (2005)	Bangladeshi villagers	370	Trust	$67.32 vs $1683	Lower offers	-

Study	Description	Number of subjects	Game Type	Comparison	Effect of Incentives	Effect on Outcome Variance
Johnson and Mislin (2011)	Meta-analysis of 162 studies	-	Trust	$0 up to $238.10	Could not assess	-
Kocher, Martinsson, and Visser (2008)	South African students	120	Public-Good	$0.25 vs $1.23	No change	-
List and Cherry (2008)	US students	310	Dictator	$20 vs $100	No change	-
List and Cherry (2000)	US students	56	Ultimatum	$20 vs $400	Less rejection	-
Munier and Zaharia (2002)	France and Romania	124	Ultimatum	$7.2 vs $360	Less rejection	-
Novakova and Flegr (2013)	Students in Czechoslovakia	524	Ultimatum	Hypothetical $1 to up hypothetical $10,000	Less rejection	-
Novakova and Flegr (2013)	Students in Czechoslovakia	524	Dictator	Hypothetical $1 to up hypothetical $10,000	Lower offers	-
Oosterbeek, Sloof, and Van de Kuilen (2004)	Meta-analysis of 37 studies	-	Ultimatum	$0.33 up to $400	Lower offers, less rejection	-
Raihani, Mace, and Lamba (2013)	Mturk Participants in US and India	1174	Dictator	$1 vs $5 vs $10	No change for US players; lower offers for Indian players	More variance
Roth et al. (1991)	Students from the US, Yugoslavia, Israel, and Japan	126	Ultimatum	$10 vs $30	No change	-
Sefton (1992)	US students	48	Ultimatum	$0 vs $5	Lower offers	-

Study	Description	Number of subjects	Game Type	Comparison	Effect of Incentives	Effect on Outcome Variance
Slonim and Roth (1998)	Slovakian players	82	Ultimatum	$1.9 vs $9.6 vs $48.4	No change in one-shot game, change in repeated game	-
Straub and Murnighan (1995)	Students	49	Ultimatum	$5 vs $100	No change	-
Tompkinson and Bethwaite (1995)	Lawyers	43	Ultimatum	Hypothetical $10 vs Hypothetical $10,000	No change	Increasing CV with higher stakes

5.3 Market Experiments

5.3.1. Why this method is relevant for informing USDA policy-making

Experimental economists use market experiments to understand how the trading environment affects market prices, trading patterns, total welfare, and the types of individuals how participate in markets. Most agricultural inputs and outputs are traded in markets. Many USDA programs and policies explicitly aim to influence the trading patterns in agricultural markets by reporting on monthly prices, production and supply estimates, and trade statistics; establishing quality grading standards for commodities; facilitating participation in domestic and international markets for disadvantaged producers; regulating market transactions for dairy and other agricultural products; etc. USDA already collects non-experimental data to inform policymaking on agricultural markets (e.g. 83 FR 11492, 83 FR 30397, 83 FR 44563, 83 FR 48794). Market experiments complement these other types of data collections by allowing for testing of causal inferences about treatment effects on market outcomes.

5.3.2. How market experiments work

Market experiments are experiments in which subjects act as buyers and sellers trading for a good. While auction experiments may also have buyers and sellers trading a good, the purpose of an auction experiment is to reveal subjects' willingness to pay/willingness to sell. In contrast, market experiments typically induce subjects' valuations in order to understand how changes in the trading environment influence the market outcomes.

Unlike real-effort tasks, these experiments rely on stated-effort in which a sellers is contracted to sell a product and is typically provided with a cost and revenue function. The good traded may be an abstract good or a real good (i.e. an endowment effect experiment). The trading environment may involve a

centralized price discovery mechanism (e.g. an auction), or rely on decentralized transactions (e.g. bilateral trading, bilateral or multilateral contracting). Experiments often induce values for buyers' willingness to pay and sellers' willingness to accept in order to study the effects of treatments on the trading environment.

5.3.3. How payment calibration affects measurement of outcomes

Error in market experiments occurs when a subject makes a decision that is inconsistent with their true preferences. When the trading environment is a double auction, error occurs when subjects make bids or sales offers inconsistent with their true valuation for the good. When the trading environment is a decentralized transaction, error occurs when (1) subjects make contract or trade offers inconsistent with their true value of the transaction, or (2) subjects accept contract or trade offers inconsistent with their true value for the transaction.

The implications of this type of error are different depending on the outcome measured. For example, consider a market experiment where a small number of subjects acting as buyers and sellers trade for a hypothetical good in a single round of bilateral contracting where buyers offer a contract to one of the sellers and the seller must decide whether to accept or reject the contract. Assume that one seller accepted a contract in error in one of the rounds of trading. A single incorrect decision to accept the contract could potentially have a larger effect on the share of contracts accepted (because this number is bounded between zero and one) than if the metric of interest was total social welfare generated from contracting (a number with no bounds).

Performance payments provide incentives for subjects to exert effort in considering their decisions within the experiment. The extent of this incentive depends on three things: the level of performance payments, the payment spread between low effort and high effort outcomes, and the extent to which subjects' individual decisions can influence common outcomes. For example, consider a double auction market experiment where buyers earn their endowment less the market-clearing price if they purchase the good and nothing if they fail to purchase the good. Assume the market clearing price occurs is equivalent to 60% of the buyer's endowment. At endowments of $1.00, the market-clearing price is $0.60, and the difference in earning between purchasing the good and not purchasing is $0.60. At endowments of $10.00, the market-clearing price is $6.00 and the incentive to succeed in purchasing the good is $6.00 as well. The $10 endowment generates stronger incentives for the buyer to want to purchase the good.

When the buyer's bids influence what the market-clearing price will be, then the strength of the incentives is also be conditional on the buyer's bidding strategy. If the buyer commits an error and bids more than his/her true willingness to pay, this would increase the market-clearing price, and decrease the buyer's earnings from purchasing the good. The earnings penalty from overbidding also scales with the level of endowments. Buyers can overbid their true preferences more for endowments of $10 than for endowments of $1, generating larger earnings penalties for overbidding with larger magnitude bidding errors.

5.3.4 Norms for incentive payments in market experiments

Market experiments typically provide subjects, divided into buyers and sellers, revenue, cost, and utility functions in terms of points or tokens. These points are then converted at the end of the experiment into currency. The studies reviewed do not suggest a particular norm for the range of incentive payments used. In general, economic experiments aim to pay a competitive wages to what a subject (typically a student) could earn for an hour worth of work outside of the laboratory.

5.3.5 Results for the impact of performance payments on outcomes

Few market experiments have analyzed the effect of performance pay on effort. Only the study by Irlenbusch and Ruchala (2008), a public good game, had enough data to calculate the coefficient of variation. The study tested payments of $0.00, $0.09, and $0.42, finding that increasing payments improved effort, and that the coefficient of variation shrunk with the increase (0.8, 0.7, 0.6, respectively for each payment level).

Other market experiments reviewed have found either positive or ambiguous effects of increased payments on effort (Cooper et al. 1999; Heyman and Ariely, 2004; Kocher, Martinsson, and Visser, 2008; Ederer and Manso, 2013).

As examples of ambiguous findings in market experiments, Cooper et al. (1999) found that sessions that provided relatively high pay led to more-frequent strategic behavior from participants in the first two-thirds of the experiment relative to sessions that provided standard pay, though the last third of the experiment high pay had no effect. Kocher, Martinsson, and Schindler (2017) assessed the effect of stake size in an experimental asset market in which traders could buy or sell shares. Half of the traders received relatively small endowments and the other half received relatively large endowments. They found that higher stakes increased the volume of trade but did not lead to asset overpricing. Ederer and Manso (2013) compared fixed-rate and performance-based (profits from business decisions), and found subjects in the fixed-rate scheme spent the least amount of time evaluating decisions and effort into recording their previous choices, but no difference in average profits between the incentive schemes.

5.4 Risk Elicitations

5.4.1. Why this method is relevant for informing USDA policy-making

There is inherently uncertainty about the future. U.S. agricultural producers face many risks, such as changes in commodity prices, unpredictable weather, and other events beyond their control. In agricultural economics, risk has long been recognized as an important driver of decisions related to production (see Antle, 1983; Hardaker, 2004; Chavas et al., 2010; Just and Pope, 2013), including pesticide use (Pannell, 1991), adoption of cover crops (Snapp et al., 2005; Schipanski et al., 2014), and adoption of crop insurance (Goodwin and Kastens, 1993; Smith and Baquet, 1996; Coble et al., 1996; Velandia et al., 2009). Economics broadly recognizes that risk enters nearly all decisions and should be accounted for to generate more accurate predictions of behavior or analyses of the effects of policy.

Economic models of risk address two components: risk exposure and risk preference. Risk exposure determines the range of possible outcomes and the probability of each outcome occurring. Risk preference is an innate characteristic of individuals and is usually assumed to be independent of the

actual degree of risk exposure. Six types of risk preferences are most commonly studied in economics experiments:

1) Known gains and probabilities (risk preference)
2) Known losses and probabilities (loss-aversion preference)
3) Unknown pay-offs or probabilities (ambiguity preference)
4) Variation in risk with time (joint risk and time preferences)
5) Variation in risk based on framing of the problem (narrow bracketing or myopia)
6) Variation based on the nature of the party that benefits (social or other-regarding preferences)

We concentrate on experiments involving the first type – risks with known gains and probabilities. When a study's methods measured multiple components of risk attitudes simultaneously, we focused on the effect of performance incentives on participants' risk attitudes.

To understand risk preferences, economists use a variety of methods, including experiments, collection of survey data on stated risk attitudes, and analysis of actual purchases of high-risk assets such as insurance policies and financial investments. Experimental studies are the most commonly used method and provide the greatest control over the design of the risk tasks and financial incentives (Charness et al., 2013). Such experiments use decisions made by subjects to characterize their individual risk attitudes and identify the distribution of risk attitudes across all subjects in the experiment.

A review of studies conducted since 1980 shows that numerous risk-elicitation methods have been used in economics experiments, and eleven methods for measuring risk-aversion have been replicated widely using diverse pools of subject. These eleven methods fall into three basic categories:

1) Paid lottery tasks (Binswanger, 1980, 1981; Holt and Laury, 2002, 2005; Eckel and Grossman, 2008; Harrison et al., 2005; Abdellaoui et al., 2007; Bruhin et al., 2010; Tanaka et al., 2010).
2) Paid investment tasks (Gneezy and Potters, 1997; Andreoni and Sprenger, 2012a, 2012b).
3) Unpaid tasks such as balloon analogues (Lejuez et al., 2002) and "bomb" risk (Crosetto and Filippin, 2013).

USDA ERS experimental research on risk attitudes has primarily relied on paid lottery tasks to elicit risk preferences (Hellerstein et al., 2013; 81 FR 63736). For this reason, we restrict our discussion to those methods.

5.4.2. How risk elicitations work

The most commonly used paid lottery methods are single-choice tasks (Binswanger, 1980; Eckel and Grossman, 2008), multiple price lists with varying probabilities (Holt and Laury, 2002; Harrison et al, 2005), certain vs risky decisions (Bruhin et al., 2010), and multiple price lists with varying payment levels (Tanaka et al., 2010). We restrict our analysis to the three most commonly employed methods: the single-choice method, the Holt and Laury multiple price list, and the certain vs. risky method.

In the Binswanger single-choice method (Binswanger, 1980, 1981), subjects choose between eight options by reporting the row containing their preferred option as shown in Table 5. Their degree of risk-aversion is inferred from the row selected. As the row number increases, the expected value of the risky choice and the spread between the potential pay-offs increase. In the example shown in the table, rows 3 and 4 have expected values of $80, rows 5 and 6 have expected values of $90, and rows 7 and 8 have expected values of $100, and a risk-neutral person would be indifferent between rows 7 and 8 since those offer the greatest expected value. A risk-averse person would choose row 1, 2, 3, 5, or 7 depending on the degree of the aversion.

Table 5: Example of a Binswanger-style Single-Choice Risk Elicitation with Responses for a Risk-Neutral Subject

Row	Risk	Preferred Option
1	100% chance of getting $50	Row 7 or 8
2	50% chance of getting $45 50% chance of getting $95	
3	50% chance of getting $40 50% chance of getting $120	
4	50% chance of getting $35 50% chance of getting $125	
5	50% chance of getting $30 50% chance of getting $150	
6	50% chance of getting $20 50% chance of getting $160	
7	50% chance of getting $10 50% chance of getting $190	
8	50% chance of getting $0 50% chance of getting $200	

The Eckel and Grossman multi-choice method (Eckel and Grossman, 2002, 2008) uses a variant of the Binswanger method to jointly study risk and loss-aversion. The menu presents five alternatives that offer different values. Subjects select only one option, providing no opportunity for subjects to make choices that are inconsistent with expected utility theory. In some versions of the method, the two rows with the greatest expected pay-offs offer lotteries that have the potential for a large winning and a negative pay-off. As in Binswanger, subjects choose the row they prefer and risk-aversion is inferred from the row selected.

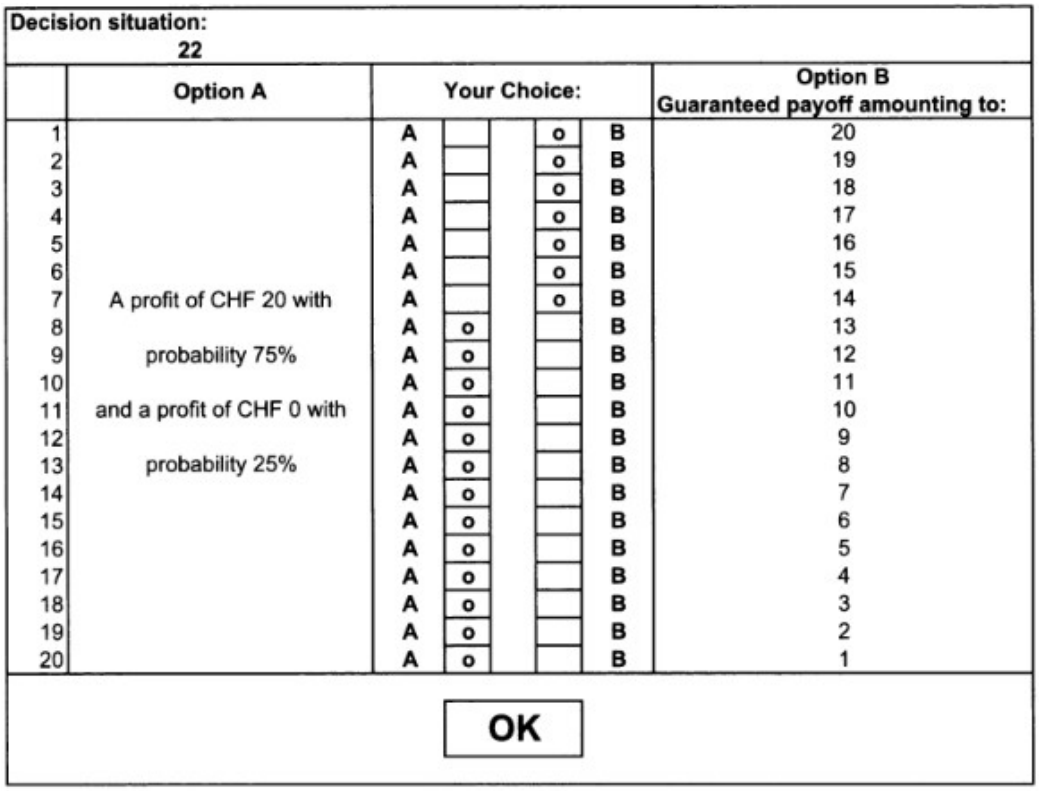
In the Holt and Laury multi-choice method (Holt and Laury, 2002, 2005), subjects make ten choices from a list like the one shown in Table 6. In each row the payments offered by option A stay the same, as do the payments offered by option B. Both options have small and large payments with some probability. Option A is always the less-risky option because the spread between pay-offs is only $0.40 (compared to a spread of $3.75 for option B), and the options are calibrated so that the greatest potential pay-off sometimes occurs in the less-risky option A and other times in the more-risky option B. The subjects record a decision (option A or B) for every row in the menu, and their degree of risk-aversion is inferred from the proportion of rows in which option A was selected. A risk-neutral person would choose the

option in each row that offers the greatest expected value (A in rows 1 through 4 and B in rows 5 through 10), and a risk-averse person would choose option A more often than a risk-neutral person.

Table 6: Example of a Holt and Laury Risk Elicitation with Responses for a Risk-Neutral Subject

Row	Option A	Option B	Preferred Option
1	10% chance of getting $2.00 90% chance of getting $1.60	10% chance of getting $3.85 90% chance of getting $0.10	A
2	20% chance of getting $2.00 80% chance of getting $1.60	20% chance of getting $3.85 80% chance of getting $0.10	A
3	30% chance of getting $2.00 70% chance of getting $1.60	30% chance of getting $3.85 70% chance of getting $0.10	A
4	40% chance of getting $2.00 60% chance of getting $1.60	40% chance of getting $3.85 60% chance of getting $0.10	A
5	50% chance of getting $2.00 50% chance of getting $1.60	50% chance of getting $3.85 50% chance of getting $0.10	B
6	60% chance of getting $2.00 40% chance of getting $1.60	60% chance of getting $3.85 40% chance of getting $0.10	B
7	70% chance of getting $2.00 30% chance of getting $1.60	70% chance of getting $3.85 30% chance of getting $0.10	B
8	80% chance of getting $2.00 20% chance of getting $1.60	80% chance of getting $3.85 20% chance of getting $0.10	B
9	90% chance of getting $2.00 10% chance of getting $1.60	90% chance of getting $3.85 10% chance of getting $0.10	B
10	100% chance of getting $2.00 0% chance of getting $1.60	100% chance of getting $3.85 0% chance of getting $0.10	B

There are several variants of the certain vs. risky task. In Bruhin et al.'s lottery method (2010), subjects made multiple choices between a certain pay-off (option B) and a risky option (option A) that would potentially pay more as shown in the sample in Figure 4. Risk-aversion is inferred from the point at which subjects switch from B to A. Other variants of the method differ in the number of rows in the menu, the number of menus to complete, and/or keep certain payment the same but vary the gamble for each row.



Decision situation: 22

	Option A	Your Choice:					Option B Guaranteed payoff amounting to:
1		A			o	B	20
2		A			o	B	19
3		A			o	B	18
4		A			o	B	17
5		A			o	B	16
6		A			o	B	15
7	A profit of CHF 20 with	A			o	B	14
8		A	o			B	13
9	probability 75%	A	o			B	12
10		A	o			B	11
11	and a profit of CHF 0 with	A	o			B	10
12		A	o			B	9
13	probability 25%	A	o			B	8
14		A	o			B	7
15		A	o			B	6
16		A	o			B	5
17		A	o			B	4
18		A	o			B	3
19		A	o			B	2
20		A	o			B	1

OK

Figure 4: Example of a Decision Menu from Bruhin et al (2010) for a Slightly Risk-Averse Subject

5.4.3. How payment calibration affects measurement of outcomes

The method used to measure participants' risk-aversion affects the estimates derived from the experiment. Csermely and Rabas (2016), for example, showed that nine similar tasks in a paid lottery scheme generated different distributions of estimated risk-aversion and that the relative ranking of subjects' attitudes toward risk was not correlated across all methods. Several other studies also tested multiple risk-elicitation methods and found low within-subject correlation across the estimates (Dave et al., 2010; Deck et al., 2010; Crosetto and Filippin, 2016).

The differences in measured risk-aversion across lottery tasks may reflect variations in the instruments' sensitivity. The Holt Laury menu includes only four payment levels. Drichoutis and Lusk (2016) pointed out that a Holt Laury menu in which the payments are fixed provides less information about an individual's risk-aversion than a menu in which the payments offered in each row vary since multiple payments provide a better estimate of the curvature of a utility function. Another possibility is that the effect of measurement error varies with the risk-elicitation task. Gillen et al. (2015) found that controlling for measurement error significantly increased observed correlation between the estimates of risk-aversion from four tasks.

Measurement error can also occur in multi-choice price-list-style elicitations when a person erroneously selects an option – A when they intended to choose B, for example. In Holt Laury multi-choice tasks, the estimates of risk-aversion are subject to measurement error in each row independently. Other approaches, such as Bruhin et al.'s choice between a certain and a risky pay-off have only one opportunity for measurement error – the switching point from B to A.

This type of error, in which the subject accidentally chooses the wrong option, has two primary effects. First, individuals will appear to be more or less risk-averse than they truly are. A risk-neutral individual who incorrectly marks B in row 4 of Table 6 will appear to be slightly risk-seeking. The experimenter would infer that the subject's risk-aversion coefficient[9] fell between –0.49 and –0.15 instead of the true range of –0.15 to 0.15. Second, groups of individuals will appear to be more or less heterogeneous in their preferences than they truly are. If subjects in a group are all equally risk-averse, measurement error will cause some of them to appear less risk-averse than they are and will increase the variance of all subjects' responses relative to the true variance of zero.

Risk estimates can be inaccurate for many reasons, including subjects' poor comprehension, lack of attention, and/or inability to evaluate the relative pay-offs and random errors. The design of the elicitation instrument is largely responsible for ensuring that the task can be understood by participants, and is not the focus of this review. For example, Levy-Garboua et al. (2012) found that presenting decisions sequentially in a Holt Laury menu results in higher rates of inconsistency than presenting all of the decisions simultaneously. For our analysis, we take the form of the menus used as given and assess, to the extent possible, how the scale of payments offered affects the quality of the responses.

Subjects must exert cognitive effort to assess the choices in each row and select one accurately. A subject who filled out the Holt Laury menu in Table 6 with random guesses for each row would expect to earn $1.99 on average. Exerting effort to evaluate each row and make a non-random choice would raise the expected payoff to $2.43 for a risk-neutral individual, a gain of $0.44. With larger payoffs, the gain from exerting effort grows. Multiplying all of the payoffs in Table 6 by 100 would raise the expected payoff from guessing randomly to $199, the expected payoff from guessing purposefully to $243, and increase the return to exerting effort to $44. Thus the size of the compensation offered should have a direct influence on how accurately subjects' choices for each row match their true preferences, and therefore the estimates derived from those choices.

In addition to affecting the attention participants give to choices, small stakes may reduce the salience of different pay-offs. Andersen et al. (2008) point out that subject may round payments to the nearest integer. While the difference between $1.75 and $1.90 may be large for the parameter estimate, subjects may not view these as meaningfully different amounts.

5.4.4 Norms for incentive payments in risk elicitations

Payments used in risk elicitations typically range from less than $1 to several hundred dollars. However the gain to exerting effort, measured as the difference in expected return from guessing randomly and guessing purposefully, rarely exceeds $5.

The conventional theory of expected utility claims that the curvature of a utility function can be measured regardless of the size of the stakes. However, when Rabin (2000) and Rabin and Thaler (2001) extrapolated estimates of risk-aversion from choices made for small stakes (in which the subjects should be nearly risk-neutral), they found that the estimates were impossibly high. A number of experiment-based studies of risk preferences conducted in the United States and several other countries (Bosch-Domenech and Silvestre, 2006; Schechter, 2007; Heinemann, 2008; Fehr-Duda et al., 2010; Andersen

[9] The coefficient of risk-aversion used in this study is the parameter α in the constant relative risk-aversion utility function $U(x) = x^{1-\alpha} / (1 - \alpha)$.

et al., 2011) and several analyses of observational data (Cohen and Einav, 2007; Sydnor, 2010; Bombardini and Trebbi, 2012) suggest that small stakes have limited ability to accurately reveal risk preferences.

It is possible that individuals have different risk attitudes for low and high stakes risks. In that case, then real-world decisions should be best reflected in lab studies using real-world stakes. Another issue associated with the range of payments used is the type of participants recruited for experiments. In the reviewed studies, payments used in students were typically less than the payments used for risk elicitations with adult subjects. Therefore, studies based on student participants may suffer more from measurement error than studies using adult subjects.

5.4.5 Results for the impact of performance payments on outcomes

Bellemare and Shearer (2010) elicited preferences using the Holt Laury method and found no difference in risk-aversion when the stakes in a safer option were doubled from 18 to 36. Faff et al. (2008) multiplied the base stakes, $1.80, by 38, 78, and 150 for real and hypothetical stakes and found that all of the risk-aversion measures were closely aligned. Conversely, Holt and Laury (2002, 2005) and Harrison et al. (2005) tested a range of scales for hypothetical and actual payments and found that scaling up the actual payments increased their measures of risk-aversion but had no effect on hypothetical payments. Reynaud and Couture (2012) used a small sample of 30 farmers with the Holt Laury and Eckel Grossman methods and found no change in risk-aversion after scaling the payment 20 times; however, their design was severely underpowered.

Using a between-subjects design, Levy-Garboua et al. (2012) assigned 120 subjects to a low-stake group (the Holt Laury values shown in Table 3) and another 120 subjects to a high-stake (10 times the low-stake values) group. They found that the high-stake group had a much lower inconsistency rate (19%) than the low-stake group (36%) ($p < 0.001$). Harrison et al. (2005) used the same scale-up of values, and while their point estimates moved in the same direction as those in Levey-Garboua et al. (2012), the difference was not as large – 10% for the high-stake group versus 16% for the low-stake group ($p = 0.25$). Bellemare and Shearer (2010), using scales of 10 and 20 times the Holt Laury values, identified inconsistency rates of 31% (for 10 times) and 26% (for 20 times) ($p = 0.55$).

Holt and Laury's (2002) study used a between-subjects design with both actual and hypothetical stakes. They also incorporated the effects of learning by having subjects complete several iterations of the task and comparing their performance on the last iteration. Their results showed small differences in consistency for the various stakes, in line with the studies previously discussed. When the stakes were relatively low (20 times the Holt Laury values), 6.6% of the subjects were inconsistent. When the stakes were higher (50 and 90 times the values), 5.5% of the subjects were inconsistent. In the treatments involving hypothetical payments, there was a small jump from 8% to 10%.

Figure 5 and Figure 6 graph the relationship between strength of incentive and coefficient of variation of the estimated risk aversion coefficient for the Holt Laury, Eckel Grossman, and Certain vs Risky tasks. The lists of each of the studies included can be found in Appendices A.2 – A.4. The Holt Laury and Certain vs. Risky tasks show a general pattern of decreasing variation in outcomes with higher incentives, although there are very few studies with high incentives for either task. The Eckel Grossman tasks has a pattern of increasing variation in outcomes with higher incentives.

It is also notable that the Holt Laury task has substantial variation around the estimated risk aversion coefficient even with high strength incentives. All but one study using the Certain vs. Risky task has a coefficient of variation less than 1, whereas eleven studies using Holt Laury menus have coefficients of variation greater than 1.5. This suggests that some tasks may be easier for subjects to understand, independent of incentive payment.

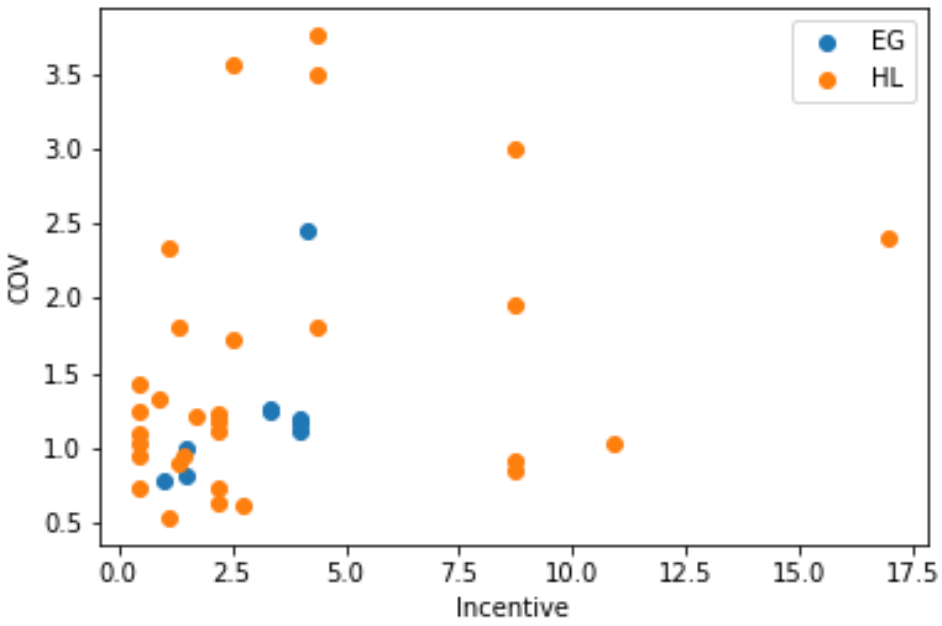


Figure 5: Effect of Incentives on Performance in Eckel Grossman and Holt Laury Tasks

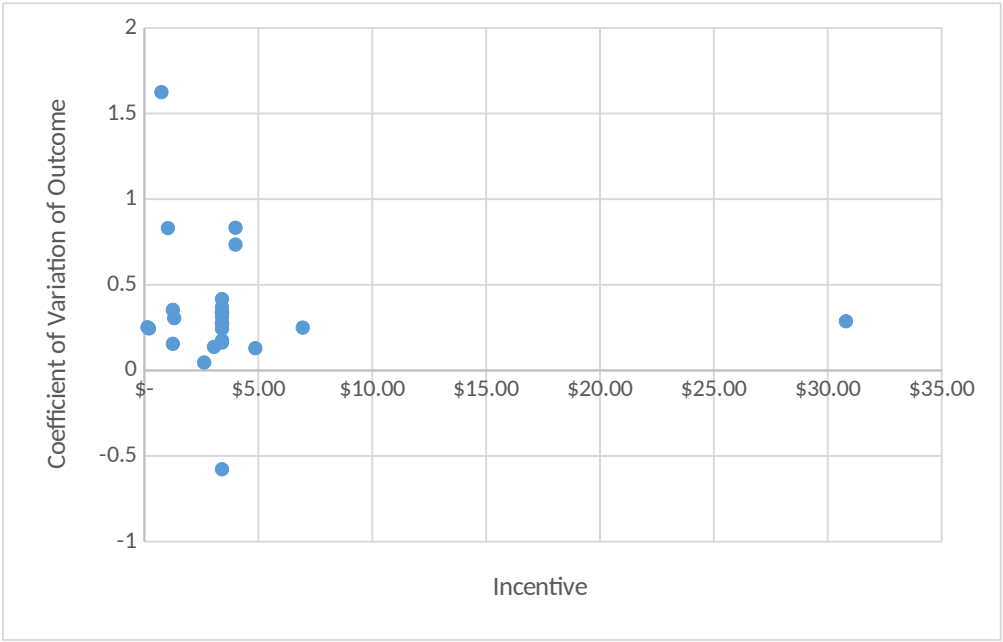


Figure 6: Effect of Incentives on Performance in Certain vs. Risky Tasks

6.0 Discussion

6.1 Synthesis of Results

We review the literature on the effects of incentive payments on outcomes in economics experiments involving real-effort tasks, auctions, multi-player games, market experiments, and risk elicitations. We also re-analyze published results to look for evidence that stronger incentives reduce the noisiness of experimental outcomes, consistent with a hypothesis that incentives improve the statistical power of experimental design.

For real effort tasks, we observe a pattern of decreased noisiness of outcomes at higher levels of piece-rate payments, quota payments, and tournament payments. We also observe patterns of decreased noisiness of outcomes at higher levels of incentives for conservation auctions, and two out of three types of risk elicitation tasks. However, the total number of studies available for these analysis and the range of payments used in those studies are both too limited to present strong evidence of a relationship between incentives and precision of measured outcomes for these methods. Sufficient data was not available to analyze the results for multi-player games or market experiments, although the few available studies suggest potential for a relationship between incentives and outcomes for these methods.

If a minimum threshold exists for incentive payments to have an effect on precision of experimental outcomes, the threshold is likely to be task specific. Piece-rate payments of less than $1 are associated with coefficients of variation in outcomes of less than 1, while quota and tournament payments above $5 sometimes results in coefficients of variation of outcomes greater than 1. Holt Laury type risk elicitations rarely have coefficients of variation of less than 1 regardless of payment levels, while nearly all studies using certain vs. risky tasks had coefficients of variation of less than 1.

6.2 Critical Gaps in the Literature

Most of the studies reviewed used relatively low values of incentive payments, and none of the methods reviewed had sufficient observations at high payment scales to be able to statistically evaluate the relationship between incentives and outcomes. Additional studies are needed, particularly studies that provide within-subject variation in incentives, covering a wide range of incentive payments, and including measures of subject effort as well as experimental outcomes.

We selected study inclusion and exclusion criteria with the aim of narrowing our analysis to well-designed and well-powered experimental studies. We also assume that the studies used incentives that were properly calibrated to the task and population sampled. None of the studies reviewed included descriptions of how payments were selected or whether pre-testing was used to verify payment calibrations were adequate to induce subjects to exert effort in the given experimental design. Additional research is needed to document norms for pre-testing of incentive payment levels, and develop recommendations for pre-testing procedures necessary to insure incentives are adequately calibrated for a given experimental design.

Our review of the literature did not identify a means of ranking the physical and cognitive burdens among the methods reviewed. Experimental research funded by USDA ERS is intended to inform policy-making, and a key issue for designing these experiments is how well the outcomes from the experiment are likely to correspond to equivalent real-world outcomes. Additional research is needed to establish a scale for ranking the physical and cognitive difficulty of different experimental methods, and to test whether methods that are physically and/or cognitively easier are more likely to have external validity for policy-making applications than methods that are more physically and/or cognitively challenging.

Our review of the literature also did not identify a standard method of measuring aptitude for different experimental methodologies in different populations. Many experimentalists ask subjects to complete test questions to verify that the subjects understand the rules of the experiment, but these questions generally do not give a view into how must effort the subjects must expend to understand the rules and how that effort varies across subjects. USDA ERS conducts research on populations with substantial heterogeneities (e.g. farmers, rural residents, US consumers). Additional research is needed to establish a scale for measuring an individual's aptitude for completing different experimental methods, and to benchmark the population average aptitude for different populations and methods of interest.

6.3 Comparison to Findings from Economics and Other Social Sciences

The experimental economics literature had previously concluded that incentive payments were necessary to induce performance. Our results also suggest that incentive payments are preferable to hypothetical (zero value) payments for most methods reviewed.

The psychology literature had previously concluded that incentive payments were unnecessary. Because experiments in psychology and other social sciences have often been low powered, our findings are also consistent with the conclusions from the experimental psychology literature: incentives are not likely to have a significant effect on the noisiness of outcomes when the experiment is too underpowered to accurately measure the outcome in the first place.

Neither literature had previously considered the question of adequate calibration of incentive payments, so our results provide a first step towards answering that question.

6.4 Framed vs Unframed Experiments

Experiments may be presented in relationship to a specific context (i.e. framed) or presented in a generic context (i.e. unframed). The literature finds mixed results on the effects of framing on decision-making within the experiment (see, for example, Cookson, 2000; Dreber et. al, 2011; Ellingsen et al, 2012; Schindler and Pfattheicher, 2017). If framing effects decision-making, there is potential that framing could also affect effect exerted in the experiment, and therefore confound our analysis of the effect of incentives on outcomes. A more robust treatment of the subject would control for the experimental framing, as well as other confounders such as number of subjects and population sampled.

References

Abdellaoui, M., Barrios, C., and Wakker, P. P. (2007). Reconciling introspective utility with revealed preference: Experimental arguments based on prospect theory. *Journal of Econometrics*, *138*(1), 356-378.

Abdellaoui, M., Bleichrodt, H., and l'Haridon, O. (2008). A tractable method to measure utility and loss aversion under prospect theory. *Journal of Risk and Uncertainty*, *36*(3), 245.

Ackert, L. F., Charupat, N., Church, B. K., & Deaves, R. (2006). An experimental examination of the house money effect in a multi-period setting. *Experimental Economics*, *9*(1), 5-16.

Adams, J. Stacy, and William B. Rosenbaum. 1962. "The Relationship of Worker Productivity to Cognitive Dissonance about Wage Inequities." *Journal of Applied Psychology* 46 (3): 161.

Adams, W. M., Brockington, D., Dyson, J., & Vira, B. (2003). Managing tragedies: understanding conflict over common pool resources. *Science*, *302*(5652), 1915-1916.

Allan, Julia, Keith A. Bender, and Ioannis Theodossiou. 2017. "Performance Pay and Stress: An Experimental Study."

Al-Ubaydli, Omar, Steffen Andersen, Uri Gneezy, and John A. List. 2015. "Carrots That Look like Sticks: Toward an Understanding of Multitasking Incentive Schemes." *Southern Economic Journal* 81 (3): 538–561.

Amir, O., & David, G. Rand, and Yaakov Kobi Gal. 2012. "Economic Games on the Internet: The Effect of $1 Stakes.". *PLoS One*, *7*(2), e31461.

Andersen, S., Ertaç, S., Gneezy, U., Hoffman, M., & List, J. A. (2011). Stakes matter in ultimatum games. *American Economic Review*, *101*(7), 3427-39.

Andersen, S., Harrison, G. W., Lau, M. I., and Rutström, E. E. (2008a). Eliciting risk and time preferences. *Econometrica*, *76*(3), 583–618.

Andersen, S., Harrison, G. W., Lau, M. I., and Elisabet Rutström, E. (2008b). Lost in state space: are preferences stable? *International Economic Review*, *49*(3), 1091-1112.

Andersen, Steffen, James Cox, Glenn Harrison, Morten Lau, Elisabet Rutstroem, and Vjollca Sadiraj. "Asset integration and attitudes to risk: theory and evidence." (2011).

Andreoni, J., and Sprenger, C. (2012a). Estimating time preferences from convex budgets. *The American Economic Review*, *102*(7), 3333-3356.

Andreoni, J., and Sprenger, C. (2012b). Risk preferences are not time preferences. *The American Economic Review*, *102*(7), 3357-3376.

Araujo, Felipe A., Erin Carbone, Lynn Conell-Price, Marli W. Dunietz, Ania Jaroszewicz, Rachel Landsman, Diego Lamé, Lise Vesterlund, Stephanie W. Wang, and Alistair J. Wilson. 2016. "The Slider Task: An Example of Restricted Inference on Incentive Effects." *Journal of the Economic Science Association* 2 (1): 1–12.

Ariely, D., U. Gneezy, George Loewenstein, and Nina Mazar. 2009. "Large Stakes and Big Mistakes." *The Review of Economic Studies* 76 (2): 451–69.

Arkes, Hal R., Robyn M. Dawes, and Caryn Christensen. 1986. "Factors Influencing the Use of a Decision Rule in a Probabilistic Task." *Organizational Behavior and Human Decision Processes* 37 (1): 93–110.

Arnold, M., J.M. Duke, and K.D. Messer. 2013. “Adverse Selection in Reverse Auctions for Environmental Services” *Land Economics*. 89(3): 387-412.

Ashraf, N., Bandiera, O., & Jack, B. K. (2014). No margin, no mission? A field experiment on incentives for public service delivery. *Journal of Public Economics*, *120*, 1-17.

Ashton, Robert H. 1990. “Pressure and Performance in Accounting Decision Settings: Paradoxical Effects of Incentives, Feedback, and Justification.” *Journal of Accounting Research*, 148–180.

Atkinson, John W., and Walter R. Reitman. 1956. “Performance as a Function of Motive Strength and Expectancy of Goal-Attainment.” *The Journal of Abnormal and Social Psychology* 53 (3): 361.

Awasthi, Vidya, and Jamie Pratt. 1990. “The Effects of Monetary Incentives on Effort and Decision Performance: The Role of Cognitive Characteristics.” *Accounting Review*, 797–811.

Bahrick, Harry P. 1954. “Incidental Learning under Two Incentive Conditions.” *Journal of Experimental Psychology* 47 (3): 170.

Bahrick, Harry P., Paul M. Fitts, and Robert E. Rankin. 1952. “Effect of Incentives upon Reactions to Peripheral Stimuli.” *Journal of Experimental Psychology* 44 (6): 400.

Bailey, Charles D., Lawrence D. Brown, and Anthony F. Cocco. 1998. “The Effects of Monetary Incentives on Worker Learning and Performance in an Assembly Task.” *Journal of Management Accounting Research*.

Bandiera, Oriana. 2007. “Contract Duration and Investment Incentives: Evidence from Land Tenancy Agreements.” *Journal of the European Economic Association* 5 (5): 953–986.

Bandiera, O., Barankay, I., & Rasul, I. (2007). Incentives for managers and inequality among workers: Evidence from a firm-level experiment. *The Quarterly Journal of Economics*, *122*(2), 729-773.

Banerjee, S., Kwasnica, A.M. and Shortle, J.S., 2015. Information and auction performance: a laboratory study of conservation auctions for spatially contiguous land management. Environmental and Resource Economics, 61(3), pp.409-431Bardsley, N., Cubitt, R., Loomes, G., Moffat, P., Starmer, C., & Sugden, R. (2010). *Experimental economics: Rethinking the rules*. Princeton University Press.

Banerjee, S. and 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), pp.1407-1426.

Baumeister, Roy F. 1984. “Choking under Pressure: Self-Consciousness and Paradoxical Effects of Incentives on Skillful Performance.” *Journal of Personality and Social Psychology* 46 (3): 610.

Bellemare, Charles, and Sabine Kröger. 2007. “On Representative Social Capital.” *European Economic Review* 51 (1): 183–202.

Bellemare, Charles, Patrick Lepage, and Bruce Shearer. 2010. “Peer Pressure, Incentives, and Gender: An Experimental Analysis of Motivation in the Workplace.” *Labour Economics* 17 (1): 276–283.

Bellemare, C., & Shearer, B. (2010). Sorting, incentives and risk preferences: Evidence from a field experiment. *Economics Letters*, *108*(3), 345-348.

Bentley, J. P., & Thacker, P. G. (2004). The influence of risk and monetary payment on the research participation decision making process. *Journal of medical ethics*, *30*(3), 293-298.

Bergum, Bruce O., and Donald J. Lehr. 1964. “Monetary Incentives and Vigilance.” *Journal of Experimental Psychology* 67 (2): 197.

Bevan, William, and Edward D. Turner. 1965. “Vigilance Performance with a Qualitative Shift in Reinforcers.” *Journal of Experimental Psychology* 70 (1): 83.

Binswanger, H. P. (1980). Attitudes toward risk: Experimental measurement in rural India. *American Journal of Agricultural Economics*, *62*(3), 395–407.

Binswanger, H. P. (1981). Attitudes toward risk: Theoretical implications of an experiment in rural India. *The Economic Journal*, *91*(364), 867–890.

Boehlje, M. (1999). Structural changes in the agricultural industries: How do we measure, analyze and understand them? *American Journal of Agricultural Economics*, *81*(5), 1028-1041.

Bombardini, Matilde, and Francesco Trebbi. "Risk aversion and expected utility theory: an experiment with large and small stakes." *Journal of the European Economic Association* 10, no. 6 (2012): 1348-1399.

Bonner, S. E., Hastie, R., Sprinkle, G. B., & Young, S. M. (2000). A review of the effects of financial incentives on performance in laboratory tasks: Implications for management accounting. *Journal of Management Accounting Research*, *12*(1), 19-64.

Bosch-Domènech, Antoni, and Joaquim Silvestre. "Reflections on gains and losses: A 2×2×7 experiment." *Journal of Risk and Uncertainty* 33, no. 3 (2006): 217-235.

Bowles, S. (2008). Policies designed for self-interested citizens may undermine" the moral sentiments": Evidence from economic experiments. *Science*, *320*(5883), 1605-1609.

Boyce, Rebecca R., Brown, Thomas C., McClelland, Gary H., Peterson, George L., Schulze, William D., 1992. An experimental examination of intrinsic values as a source of the WTA–WTP disparity. American Economic Review 82 (5), 1366–1373.

Bracha, Anat, Uri Gneezy, and George Loewenstein. 2015. "Relative Pay and Labor Supply." *Journal of Labor Economics* 33 (2): 297–315.

Briggeman, B. C., & Lusk, J. L. (2010). Preferences for fairness and equity in the food system. *European Review of Agricultural Economics*, *38*(1), 1-29.

Bruhin, A., Fehr-Duda, H., and Epper, T. (2010). Risk and rationality: Uncovering heterogeneity in probability distortion. *Econometrica*, 1375–1412.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data?. *Perspectives on psychological science*, *6*(1), 3-5.

Burchett, Richard, and John Willoughby. 2004. "Work Productivity When Knowledge of Different Reward Systems Varies: Report from an Economic Experiment." *Journal of Economic Psychology* 25 (5): 591–600.

Byram, Stephanie J. 1997. "Cognitive and Motivational Factors Influencing Time Prediction." *Journal of Experimental Psychology: Applied* 3 (3): 216.

Cadsby, C. B., Song, F., & Tapon, F. (2007). Sorting and incentive effects of pay for performance: An experimental investigation. *Academy of management journal*, *50*(2), 387-405.

Camerer, C. F., and Hogarth, R. M. (1999). The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework. *Journal of Risk and Uncertainty*, *19*(1), 7-42.

Camerer, Colin F., Robin M. Hogarth, David V. Budescu, and Catherine Eckel. 1999. "The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework." In *Elicitation of Preferences*, 7–48. Springer.

Cameron, L. A. (1999). Raising the stakes in the ultimatum game: Experimental evidence from Indonesia. *Economic Inquiry*, *37*(1), 47-59.

Campbell, Donald J. 1984. “THE EFFECTS OF GOAL-CONTINGENT PAYMENT ON THE PERFORMANCE OF A COMPLEX TASK 1.” *Personnel Psychology* 37 (1): 23–40.

Carpenter, Jeffrey, and Erick Gong. 2016. “Motivating Agents: How Much Does the Mission Matter?” *Journal of Labor Economics* 34 (1): 211–36.

Carpenter, Jeffrey, Peter Hans Matthews, and John Schirm. 2010. “Tournaments and Office Politics: Evidence from a Real Effort Experiment.” *The American Economic Review* 100 (1): 504–17.

Carpenter, Jeffrey, Eric Verhoogen, and Stephen Burks. 2005. “The Effect of Stakes in Distribution Experiments.” *Economic Letters* 86 (3): 393–98.

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

Cason, Timothy N., William A. Masters, and Roman M. Sheremeta. 2010. “Entry into Winner-Take-All and Proportional-Prize Contests: An Experimental Study.” *Journal of Public Economics* 94 (9–10): 604–611.

Cassar and Meier. (2018). Nonmonetary Incentives and the Implications of Work as a Source of Meaning. *Journal of Economic Perspectives*, 32(3), 215-238.

Chaddad, F. R., & Cook, M. L. (2004). Understanding new cooperative models: an ownership–control rights typology. *Applied Economic Perspectives and Policy*, *26*(3), 348-360.

Chang, J.B., J.L. Lusk, and F.B. Norwood 2009. “How Closely Do Hypothetical Surveys and Laboratory Experiments Predict Field Behavior?,” *American Journal of Agricultural Economics 91*(2): 518-534.

Charness, Gary, Ramon Cobo-Reyes, and Angela Sanchez. 2016. “The Effect of Charitable Giving on Workers’ Performance: Experimental Evidence.” *Journal of Economic Behavior & Organization* 131: 61–74.

Charness, G., and Gneezy, U. (2012). Strong Evidence for Gender Differences in Risk Taking. *Journal of Economic Behavior and Organization*, *83*(1), 50–58 https://doi.org/10.1016/j.jebo.2011.06.007.

Charness, G., Gneezy, U., and Imas, A. (2013). Experimental methods: Eliciting risk preferences. *Journal of Economic Behavior and Organization*, *87*, 43–51.

Cherry, T. L., Frykblom, P., & Shogren, J. F. (2002). Hardnose the dictator. *American Economic Review*, *92*(4), 1218-1221.

Chung, Kae H., and W. Dean Vickery. 1976. “Relative Effectiveness and Joint Effects of Three Selected Reinforcements in a Repetitive Task Situation.” *Organizational Behavior and Human Performance* 16 (1): 114–142.

Cohen, Alma, and Liran Einav. “Estimating risk preferences from deductible choice.” *American Economic Review* 97, no. 3 (2007): 745-788.

Conte, M.N. and 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), pp.571-591.

Cookson, R. (2000). Framing effects in public goods experiments. *Experimental Economics*, *3*(1), 55-79.

Cooper, David J, John H Kagel, Wei Lo, and Qing Liang Gu. 1999. “Gaming against Managers in Incentive Systems: Experimental Results with Chinese Students and Chinese Managers.” *The American Economic Review* 89 (4): 781–804.

Corgnet, B., Hernán-González, R., Kujal, P., & Porter, D. (2014). The effect of earned versus house money on price bubble formation in experimental asset markets. *Review of Finance*, *19*(4), 1455-1488.

Corgnet, Brice, Joaquín Gómez-Miñambres, and Roberto Hernán-Gonzalez. 2015. "Goal Setting and Monetary Incentives: When Large Stakes Are Not Enough." *Management Science* 61 (12): 2926–2944.

Crosetto, P., and Filippin, A. (2013). The "bomb" risk elicitation task. *Journal of Risk and Uncertainty*, 47(1), 31–65 https://doi.org/10.1007/s11166-013-9170-z.

Crosetto, P., and Filippin, A. (2016). A theoretical and experimental appraisal of four risk elicitation methods. *Experimental Economics*, *19*(3), 613-641.

Croson, R. (2005). The method of experimental economics. *International Negotiation*, *10*(1), 131-148.

Csermely, T., and Rabas, A. (2016). How to reveal people's preferences: Comparing time consistency and predictive power of multiple price list risk elicitation methods. *Journal of Risk and Uncertainty*, *53*(2-3), 107-136.

Cummings, R. G., Harrison, G. W., & Rutström, E. E. (1995). Homegrown values and hypothetical surveys: is the dichotomous choice approach incentive-compatible?. *The American Economic Review*, *85*(1), 260-266.

Dalton, Patricio, Victor Gonzalez Jimenez, and Charles Noussair. 2016. "SELF-CHOSEN GOALS: INCENTIVES AND GENDER DIFFERENCES." *CentER Discussion Paper*

Dave, C., Eckel, C. C., Johnson, C. A., and Rojas, C. (2010). Eliciting risk preferences: When is simple better? *Journal of Risk and Uncertainty*, *41*(3), 219-243.

Dechenaux, E., Kovenock, D. and Sheremeta, R.M., 2015. A survey of experimental research on contests, all-pay auctions and tournaments. Experimental Economics, 18(4), pp.609-669.

Deck, C., Lee, J., and Reyes, J. (2010). Personality and the consistency of risk taking behavior: Experimental evidence. *University of Arkansas, Department of Economics*, 1-10.

DellaVigna, Stefano, and Devin Pope. 2017. "What Motivates Effort? Evidence and Expert Forecasts." *The Review of Economic Studies* 85 (2): 1029–1069.

Diamond, P.A., and J.A. Hausman 1994. "Contingent Valuation: Is Some Number Better Than No Number?," *The Journal of economic perspectives 8*(4): 45-64.

Diekmann, A. (2004). The power of reciprocity: Fairness, reciprocity, and stakes in variants of the dictator game. *Journal of conflict resolution*, *48*(4), 487-505.

Dillard, Jesse F., and Joseph G. Fisher. 1990. "Compensation Schemes, Skill Level, and Task Performance: An Experimental Examination." *Decision Sciences* 21 (1): 121–137.

Dohmen, T., & Falk, A. (2011). Performance pay and multidimensional sorting: Productivity, preferences, and gender. *American Economic Review*, *101*(2), 556-90.

Dornbush, Rhea L. 1965. "Motivation and Positional Cues in Incidental Learning." *Perceptual and Motor Skills* 20 (3): 709–14.

Dreber, A., Ellingsen, T., Johannesson, M., & Rand, D. G. (2013). Do people care about social context? Framing effects in dictator games. *Experimental Economics*, *16*(3), 349-371.

Drichoutis, A. C., and Lusk, J. L. (2016). What can multiple price lists really tell us about risk preferences? *Journal of Risk and Uncertainty*, *53*(2-3), 89-106.

Duke, J., K.D. Messer, L. Lynch, and T. Li. 2017. "The Effect of Information on Discriminatory-Price and Uniform-Price Reverse Auction Efficiency: An Experimental Economics Study of the Purchase of Ecosystem Services." *Strategic Behavior and the Environment*. 7(1-2): 41-71.

Eckartz, Katharina, Oliver Kirchkamp, and Daniel Schunk. 2012. "How Do Incentives Affect Creativity?"

Eckel, C. C., and Grossman, P. J. (2002). Sex differences and statistical stereotyping in attitudes toward financial risk. *Evolution and human behavior*, *23*(4), 281-295.

Eckel, C. C., and Grossman, P. J. (2008). Forecasting risk attitudes: An experimental study using actual and forecast gamble choices. *Journal of Economic Behavior and Organization*, *68*(1), 1–17.

Ederer, Florian, and Gustavo Manso. 2013. "Is Pay-for-Performance Detrimental to Innovation?" *Management Science* 59 (7).

El-Gamal, Mahmoud A., and David M. Grether. 1995. "Are People Bayesian? Uncovering Behavioral Strategies." *Journal of the American Statistical Association* 90 (432): 1137–1145.

Ellingsen, T., Johannesson, M., Mollerstrom, J., & Munkhammar, S. (2012). Social framing effects: Preferences or beliefs?. *Games and Economic Behavior*, *76*(1), 117-130.

Ellison, B., Lusk, J. L., & Briggeman, B. (2010). Other-regarding behavior and taxpayer preferences for farm policy. *The BE Journal of Economic Analysis & Policy*, *10*(1).

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

Enzle, Michael E., and June M. Ross. 1978. "Increasing and Decreasing Intrinsic Interest with Contingent Rewards: A Test of Cognitive Evaluation Theory." *Journal of Experimental Social Psychology* 14 (6): 588–597.

Erez, Miriam, Daniel Gopher, and Nira Arzi. 1990. "Effects of Goal Difficulty, Self-Set Goals, and Monetary Rewards on Dual Task Performance." *Organizational Behavior and Human Decision Processes* 47 (2): 247–269.

Eriksson, Tor Viking, Anders Poulsen, and Marie Claire Villeval. 2008. "Feedback, Incentives and Peer Effects: Experimental Evidence." In *Tournaments, Contests, and Relative Performance Evaluation*.

Erkal, Nisvan, Lata Gangadharan, and Boon Han Koh. 2018. "Monetary and Non-Monetary Incentives in Real-Effort Tournaments." *European Economic Review* 101: 528–545.

Farh, Jiing-Lih, Rodger W. Griffeth, and David B. Balkin. 1991. "Effects of Choice of Pay Plans on Satisfaction, Goal Setting, and Performance." *Journal of Organizational Behavior* 12 (1): 55–62.

Farr, James L. 1976. "Task Characteristics, Reward Contingency, and Intrinsic Motivation." *Organizational Behavior and Human Performance* 16 (2): 294–307.

Farr, James L., Robert J. Vance, and Robert M. McIntyre. 1977. "Further Examinations of the Relationship between Reward Contingency and Intrinsic Motivation." *Organizational Behavior and Human Performance* 20 (1): 31–53.

Fatseas, Victor A., and Mark K. Hirst. 1992. "Incentive Effects of Assigned Goals and Compensation Schemes on Budgetary Performance." *Accounting and Business Research* 22 (88): 347–55.

Fehr, E., & Goette, L. (2007). Do workers work more if wages are high? Evidence from a randomized field experiment. *American Economic Review*, *97*(1), 298-317.

Fehrenbacher, Dennis D., and Burkhard Pedell. 2012. "Disentangling Incentive Effects from Sorting Effects: An Experimental Real-Effort Investigation." *The Wharton School, University of Pennsylvania. Risk Management and Decision Process Center, Working Paper# 2012* 8.

Fehr-Duda, Helga, Adrian Bruhin, Thomas Epper, and Renate Schubert. "Rationality on the rise: Why relative risk aversion increases with stake size." *Journal of Risk and Uncertainty* 40, no. 2 (2010): 147-180.

Ferraro, P.J., 2008. Asymmetric information and contract design for payments for environmental services. Ecological economics, 65(4), pp.810-821.

Fest, Sebastian, Ola Kvaloy, Petra Nieken, and Anja Schöttner. 2019. "Motivation and Incentives in an Online Labor Market."

Fifer, S., J. Rose, and S. Greaves 2014. “Hypothetical Bias in Stated Choice Experiments: Is It a Problem? And If So, How Do We Deal with It?,” *Transportation research part A: policy and practice 61*: 164-177.

Fooks, J.R., N. Higgins, K.D. Messer, J.M. Duke, D. Hellerstein, and L. Lynch. 2016. “Conserving Spatially Explicit Benefits in Ecosystem Service Markets: Experimental Tests of Network Bonuses and Spatial Targeting” *American Journal of Agricultural Economics.* 98(2): 468-488.

Fooks, J., K.D. Messer, and J. Duke. 2015. “Dynamic Entry, Reverse Auctions, and the Purchase of Environmental Services.” *Land Economics.* 91(1): 57-75.

Fooks, J. K.D. Messer, J. Duke, J. Johnson, T. Li, G. Parsons, 2017. “Tourist Viewshed Externalities and Wind Energy Production.” *Agricultural and Resource Economics Review.* 46(2): 224-241.

Forsythe, R., Horowitz, J. L., Savin, N. E., & Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic behavior*, *6*(3), 347-369.

Freeman, Richard B., and Alexander M. Gelber. 2010. “Prize Structure and Information in Tournaments: Experimental Evidence.” *American Economic Journal: Applied Economics* 2 (1): 149–64.

Frey, Bruno S., and Felix Oberholzer-Gee. 1997. “The Cost of Price Incentives: An Empirical Analysis of Motivation Crowding-Out.” *The American Economic Review* 87 (4): 746–755.

Friedl, Andreas, Levent Neyse, and Ulrich Schmidt. 2018. “Payment Scheme Changes and Effort Adjustment: The Role of 2D: 4D Digit Ratio.” *Journal of Behavioral and Experimental Economics* 72: 86–94.

Frisch, Carol J., and Alyce M. Dickinson. 1990. “Work Productivity as a Function of the Percentage of Monetary Incentives to Base Pay.” *Journal of Organizational Behavior Management* 11 (1): 13–34.

Gächter, Simon, Lingbo Huang, and Martin Sefton. 2016. “Combining ‘Real Effort’ with Induced Effort Costs: The Ball-Catching Task.” *Experimental Economics* 19 (4): 687–712.

Georgellis, Y., Iossa, E., & Tabvuma, V. (2010). Crowding out intrinsic motivation in the public sector. *Journal of Public Administration Research and Theory*, *21*(3), 473-493.

Gigerenzer, Gerd, Ulrich Hoffrage, and Heinz Kleinbölting. 1991. “Probabilistic Mental Models: A Brunswikian Theory of Confidence.” *Psychological Review* 98 (4): 506.

Gillis, M. T., & Hettler, P. L. (2007). Hypothetical and real incentives in the ultimatum game and Andreoni's public goods game: An experimental study. *Eastern Economic Journal*, *33*(4), 491-510.

Gillen, B., Snowberg, E., and Yariv, L. (2015). *Experimenting with measurement error: techniques with applications to the Caltech cohort study* (No. w21517). National Bureau of Economic Research.

Glucksberg, Sam. 1962. “The Influence of Strength of Drive on Functional Fixedness and Perceptual Recognition.” *Journal of Experimental Psychology* 63 (1): 36.

Gneezy, U., M. Niederle, and A. Rustichini. 2003. “Performance in Competitive Environments: Gender Differences.” *The Quarterly Journal of Economics* 118 (3): 1049–74.

Gneezy, U., and Potters, J. (1997). An experiment on risk taking and evaluation periods. *The Quarterly Journal of Economics*, *112*(2), 631–645.

Gneezy, U., Meier, S., & Rey-Biel, P. (2011). When and why incentives (don't) work to modify behavior. *Journal of Economic Perspectives*, *25*(4), 191-210.

Gneezy, Uri, and Aldo Rustichini. 2000. “PAY ENOUGH OR DON’T PAY AT ALL.” *Quarterly Journal of Economics* 115 (3): 791–810.

Goerg, Sebastian J., Sebastian Kube, and Jonas Radbruch. 2017. “The Effectiveness of Incentive Schemes in the Presence of Implicit Effort Costs.”

Gracia, A., M.L. Loureiro, and R.M. Nayga Jr 2011. “Are Valuations from Nonhypothetical Choice Experiments Different from Those of Experimental Auctions?,” *American Journal of Agricultural Economics 93*(5): 1358-1373.

Greiner, Ben, Axel Ockenfels, and Peter Werner. 2011. “Wage Transparency and Performance: A Real-Effort Experiment.” *Economics Letters* 111 (3): 236–238.

Grether, David M. 1980. “Bayes Rule as a Descriptive Model: The Representativeness Heuristic.” *The Quarterly Journal of Economics* 95 (3): 537–557.

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of economic behavior & organization*, *3*(4), 367-388.

Hamilton, Barton H., Jack A. Nickerson, and Hideo Owan. 2003. “Team Incentives and Worker Heterogeneity: An Empirical Analysis of the Impact of Teams on Productivity and Participation.” *Journal of Political Economy* 111 (3): 465–497.

Hamner, W.Clay, and Lawrence W. Foster. 1975. “Are Intrinsic and Extrinsic Rewards Additive: A Test of Deci’s Cognitive Evaluation Theory of Task Motivation.” *Organizational Behavior and Human Performance* 14 (3): 398–415.

Hanemann, W.M. 1994. “Valuing the Environment through Contingent Valuation,” *The journal of economic perspectives 8*(4): 19-43.

Harley, Willard. 1965. “The Effect of Monetary Incentive in Paired Associate Learning Using an Absolute Method.” *Psychonomic Science* 3 (1–12): 141–142.

Harley, Willard F. 1968. “Delay of Incentive Cues in Paired-Associate Learning and Its Effect on Organizing Responses.” *Journal of Verbal Learning and Verbal Behavior* 7 (5): 924–929.

Harrison, G. W. (1989). Theory and misbehavior of first-price auctions. *The American Economic Review*, 749-762.

Harrison, G. W., Igel Lau, M., Rutström, E. E., and Sullivan, M. B. (2005). Eliciting risk and time preferences using field experiments: Some methodological issues. In *Field experiments in economics* (pp. 125-218). Emerald Group Publishing Limited.

Harrison, G. W., Lau, M. I., and Rutström, E. E. (2007). Estimating risk attitudes in Denmark: A field experiment. *The Scandinavian Journal of Economics*, *109*(2), 341-368.

Harrison, G. W., Lau, M. I., & Rutström, E. E. (2009). Risk attitudes, randomization to treatment, and self-selection into experiments. *Journal of Economic Behavior & Organization*, *70*(3), 498-507.

Heinemann, Frank. “Measuring risk aversion and the wealth effect.” In *Risk aversion in experiments*, pp. 293-313. Emerald Group Publishing Limited, 2008.

Hell, Wolfgang, Gerd Gigerenzer, Siegfried Gauggel, Maria Mall, and Michael Müller. 1988. “Hindsight Bias: An Interaction of Automatic and Motivational Factors?” *Memory & Cognition* 16 (6): 533–538.

Hellerstein, D., Higgins, N., & Horowitz, J. (2013). The predictive power of risk preference measures for farming decisions. *European Review of Agricultural Economics*, *40*(5), 807-833.

Hellerstein, D., Higgins, N.A. and Roberts, M., 2015. Options for improving conservation programs: Insights from auction theory and economic experiments. Amber Waves, February.

Hennig-Schmidt, Heike, Abdolkarim Sadrieh, and Bettina Rockenbach. 2010. “IN SEARCH OF WORKERS’ REAL EFFORT RECIPROCITY—A FIELD AND A LABORATORY EXPERIMENT.” *Journal of the European Economic Association* 8 (4): 817–37.

Henrich, J. (2000). Does culture matter in economic behavior? Ultimatum game bargaining among the Machiguenga of the Peruvian Amazon. *American Economic Review*, *90*(4), 973-979.
Henry, Rebecca A., and Janet A. Sniezek. 1993. "Situational Factors Affecting Judgments of Future Performance." *Organizational Behavior and Human Decision Processes* 54 (1): 104–132.
Henry, Rebecca A., and Oriel J. Strickland. 1994. "Performance Self-Predictions: The Impact of Expectancy Strength and Incentives." *Journal of Applied Social Psychology* 24 (12): 1056–1069.
Hensher, D.A. 1997. "Behavioral Value of Travel Time Savings in Personal and Commercial Automobile Travel," *Greene, Jones and Delucchi (1997)*: 245-279.
Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences*, *24*(3), 383-403.
Heyman, James, and Dan Ariely. 2004. "Effort for Payment a Tale of Two Markets." *Psychological Science* 15 (11): 787–93.
Heyman, J. E., Orhun, Y., & Ariely, D. (2004). Auction fever: The effect of opponents and quasi-endowment on product valuations. *Journal of interactive Marketing*, *18*(4), 7-21.
Hoffman, V, J. Fooks, and K.D. Messer. 2014. "Measuring and Mitigating HIV Stigma: A Framed Field Experiment" *Economic Development and Cultural Change.* 62(4): 701-726.
Hoffman, E., McCabe, K. A., & Smith, V. L. (1996). On expectations and the monetary stakes in ultimatum games. *International Journal of Game Theory*, *25*(3), 289-301.
Hogarth, Robin M., Brian J. Gibbs, Craig R. McKenzie, and Margaret A. Marquis. 1991. "Learning from Feedback: Exactingness and Incentives." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 17 (4): 734.
Holm, H., & Nystedt, P. (2008). Trust in surveys and games–A methodological contribution on the influence of money and location. *Journal of Economic Psychology*, *29*(4), 522-542.
Holt, Charles A., and Susan K. Laury. "Risk aversion and incentive effects." *American Economic Review* 92, no. 5 (2002): 1644-1655.
Holt, Charles A., and Susan K. Laury. "Risk aversion and incentive effects: New data without order effects." *American Economic Review* 95, no. 3 (2005): 902-904.
Hubbard, Thomas N. 2003. "Information, Decisions, and Productivity: On-Board Computers and Capacity Utilization in Trucking." *American Economic Review* 93 (4): 1328–1353.
Huber, Vandra L. 1985. "Comparison of Monetary Reinforcers and Goal Setting as Learning Incentives." *Psychological Reports* 56 (1): 223–235.
Iftekhar, M.S. and Tisdell, J.G., 2014. Wildlife corridor market design: An experimental analysis of the impact of project selection criteria and bidding flexibility. Ecological economics, 104, pp.50-60.
Irlenbusch, Bernd, and Gabriele K. Ruchala. 2008. "Relative Rewards within Team-Based Compensation." *Labour Economics* 15 (2): 141–167.
Irlenbusch, Bernd, and Dirk Sliwka. 2005. "Incentives, Decision Frames, and Motivation Crowding out-an Experimental Investigation."
Jacquemet, N., Joule, R. V., Luchini, S., & Shogren, J. F. (2009). Earned wealth, engaged bidders? Evidence from a second-price auction. *Economics Letters*, *105*(1), 36-38.
Jenkins Jr, G. Douglas, Atul Mitra, Nina Gupta, and Jason D. Shaw. 1998. "Are Financial Incentives Related to Performance? A Meta-Analytic Review of Empirical Research." *Journal of Applied Psychology* 83 (5): 777.
Johansson-Stenman, O., Mahmud, M., & Martinsson, P. (2005). Does stake size matter in trust games?. *Economics Letters*, *88*(3), 365-369.

Johansson-Stenman, O., and H. Svedsäter 2008. “Measuring Hypothetical Bias in Choice Experiments: The Importance of Cognitive Consistency,” *The BE Journal of Economic Analysis and Policy 8*(1).

Johnson, Douglas A., and Alyce M. Dickinson. 2010. “Employee-of-the-Month Programs: Do They Really Work?” *Journal of Organizational Behavior Management* 30 (4): 308–24.

Johnson, N. D., & Mislin, A. A. (2011). Trust games: A meta-analysis. *Journal of Economic Psychology*, *32*(5), 865-889.

Jorgenson, Dale O., and Marvin D. Dunnette. 1973. “Effects of the Manipulation of a Performance-Reward Contingency on Behavior in a Simulated Work Setting.” *Journal of Applied Psychology* 57 (3): 271.

Kachelmeier, Steven J., Bernhard E. Reichert, and Michael G. Williamson. 2008. “Measuring and Motivating Quantity, Creativity, or Both.” *Journal of Accounting Research* 46 (2): 341–373.

Kagel, J.H. and Levin, D., 2011. Auctions: a survey of experimental research, 1995–2010. Handbook of experimental economics, 2.

Kagel, J. H., & Roth, A. E. (Eds.). (2016). *The handbook of experimental economics, volume 2: the handbook of experimental economics*. Princeton university press.

Kausler, Donald H., and E. Philip Trapp. 1962. “Effects of Incentive-Set and Task Variables on Relevant and Irrelevant Learning in Serial Verbal Learning.” *Psychological Reports* 10 (2): 451–457.

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

Kecinski, M., K.D. Messer, L. Knapp, and Y. Shirazi. 2017. “Consumer Preferences for Oyster Attributes: Field Experiments on Brand, Locality, and Growing Method.” *Agricultural and Resource Economics Review*. 46(2): 315-337.

Keisner, D.K., K.D. Messer, W.D. Schulze, and H. Zarghamee. 2013. “Testing Social Preferences for an Economic ‘Bad’: An Artefactual Field Experiment.” *Scandinavian Journal of Economics* 115(1): 27–61.

Kernoff, Phyllis, Bernard Weiner, and Myrna Morrison. 1966. “Affect and Short-Term Retention.” *Psychonomic Science* 4 (1): 75–76.

Kocher, Martin, Peter Martinsson, and Martine Visser. 2008. “Does Stake Size Matter for Cooperation and Punishment?” *Economic Letters* 99 (3): 508–11.

Korn, J. H., & Hogan, K. (1992). Effect of incentives and aversiveness of treatment on willingness to participate in research. *Teaching of Psychology*, *19*(1), 21-24.

Latacz-Lohmann, U., & Van der Hamsvoort, C. (1997). Auctioning conservation contracts: a theoretical analysis and an application. *American Journal of Agricultural Economics*, *79*(2), 407-418.

Lazear, Edward P. 2000. “Performance Pay and Productivity.” *American Economic Review* 90 (5): 1346–1361.

Lee, Thomas W., Edwin A. Locke, and Soo H. Phan. 1997. “Explaining the Assigned Goal-Incentive Interaction: The Role of Self-Efficacy and Personal Goals.” *Journal of Management* 23 (4): 541–559.

Lejuez, C. W., Aklin, W. M., Zvolensky, M. J., and Pedulla, C. M. (2003). Evaluation of the Balloon Analogue Risk Task (BART) as a predictor of adolescent real-world risk-taking behaviours. *Journal of Adolescence*, *26*(4), 475–479 https://doi.org/10.1016/S0140-1971(03)00036-8.

Levitt, S. D., List, J. A., Neckermann, S., & Sadoff, S. (2016). The behavioralist goes to school: Leveraging behavioral economics to improve educational performance. *American Economic Journal: Economic Policy*, *8*(4), 183-219.

Lévy-Garboua, L., Maafi, H., Masclet, D., & Terracol, A. (2012). Risk aversion and framing effects. *Experimental Economics*, *15*(1), 128-144.

Libby, Robert, and Marlys Gascho Lipe. 1992. "Incentives, Effort, and the Cognitive Processes Involved in Accounting-Related Judgments." *Journal of Accounting Research*, 249–273.

Lim, K. H., Hu, W., Maynard, L. J., & Goddard, E. (2013). US consumers' preference and willingness to pay for country-of-origin-labeled beef steak and food safety enhancements. *Canadian Journal of Agricultural Economics/Revue Canadienne d'agroeconomie*, *61*(1), 93-118.

List, J. A., & Cherry, T. L. (2000). Learning to accept in ultimatum games: Evidence from an experimental design that generates low offers. *Experimental Economics*, *3*(1), 11-29.

List, J. A., & Cherry, T. L. (2008). Examining the role of fairness in high stakes allocation decisions. *Journal of Economic Behavior & Organization*, *65*(1), 1-8.

List, J.A., and C.A. Gallet 2001. "What Experimental Protocol Influence Disparities between Actual and Hypothetical Stated Values?," *Environmental and Resource Economics 20*(3): 241-254.

Little, J., and R. Berrens 2004. "Explaining Disparities between Actual and Hypothetical Stated Values: Further Investigation Using Meta-Analysis," *Economics Bulletin 3*(6): 1-13.

Locke, Edwin A., Judith F. Bryan, and Lorne M. Kendall. 1968. "Goals and Intentions as Mediators of the Effects of Monetary Incentives on Behavior." *Journal of Applied Psychology* 52 (2): 104.

London, Manuel, and Greg R. Oldham. 1976. "Effects of Varying Goal Types and Incentive Systems on Performance and Satisfaction." *Academy of Management Journal* 19 (4): 537–546.

London, Manuel, and Greg R. Oldham. 1977. "A Comparison of Group and Individual Incentive Plans." *Academy of Management Journal* 20 (1): 34–41.

Loureiro, M. L., & Umberger, W. J. (2007). A choice experiment model for beef: What US consumer responses tell us about relative preferences for food safety, country-of-origin labeling and traceability. *Food policy*, *32*(4), 496-514.

Malmendier, U., & Lee, Y. H. (2011). The bidder's curse. *American Economic Review*, *101*(2), 749-87.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior research methods*, *44*(1), 1-23.

McGraw, Kenneth O., and John C. McCullers. 1979. "Evidence of a Detrimental Effect of Extrinsic Incentives on Breaking a Mental Set." *Journal of Experimental Social Psychology* 15 (3): 285–94.

McNamara, H. J., and R. I. Fisch. 1964. "Effect of High and Low Motivation on Two Aspects of Attention." *Perceptual and Motor Skills* 19 (2): 571–578.

Mellström, C., & Johannesson, M. (2008). Crowding out in blood donation: Was Titmuss right? *Journal of the European Economic Association*, *6*(4), 845-863.

Merlo, A., & Schotter, A. (1992). Theory and misbehavior of first-price auctions: Comment. *The American Economic Review*, *82*(5), 1413-1425.

Messer, K.D., J. Duke, and L. Lynch. 2014. "Applying Experimental Economics to Land Economics: Public Information and Auction Efficiency in Land Preservation Markets." in the Oxford Handbook of Land Economics. J. Duke and J. Wu editors. Oxford Press.

Messer K.D., J. Duke, L. Lynch, and T. Li. 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., G.L. Poe, D. Rondeau, W.D. Schulze and C. Vossler. 2010. “Social Preferences and Voting: An Exploration Using a Novel Preference Revealing Mechanism.” *Journal of Public Economics* 94(3-4): 308-317.

Montague, William E., and Carl E. Webber. 1965. “Effects of Knowledge of Results and Differential Monetary Reward on Six Uninterrupted Hours of Monitoring.” *Human Factors* 7 (2): 173–180.

Munier, B., & Zaharia, C. (2002). High stakes and acceptance behavior in ultimatum bargaining. *Theory and Decision*, *53*(3), 187-207.

Murphy, J.J., P.G. Allen, T.H. Stevens, and D. Weatherhead 2005. “A Meta-Analysis of Hypothetical Bias in Stated Preference Valuation,” *Environmental and Resource Economics 30*(3): 313-325.

Murphy, J. J., Stevens, T. H., & Yadav, L. (2010). A comparison of induced value and home-grown value experiments to test for hypothetical bias in contingent valuation. *Environmental and Resource Economics*, *47*(1), 111-123.

Nilsson, Lars-Göran. 1987. “Motivated Memory: Dissociation between Performance Data and Subjective Reports.” *Psychological Research* 49 (2–3): 183–188.

Novakova, J., & Flegr, J. (2013). How much is our fairness worth? The effect of raising stakes on offers by proposers and minimum acceptable offers in dictator and ultimatum games. *PloS one*, *8*(4), e60966.

Oosterbeek, H., Sloof, R., & Van De Kuilen, G. (2004). Cultural differences in ultimatum game experiments: Evidence from a meta-analysis. *Experimental economics*, *7*(2), 171-188.

Österberg, P., & Nilsson, J. (2009). Members' perception of their participation in the governance of cooperatives: the key to trust and commitment in agricultural cooperatives. *Agribusiness: An International Journal*, *25*(2), 181-197.

Ostrom, E. (2010). Polycentric systems for coping with collective action and global environmental change. *Global Environmental Change*, *20*(4), 550-557.

Paarsch, Harry J., and Bruce Shearer. 2000. “Piece Rates, Fixed Wages, and Incentive Effects: Statistical Evidence from Payroll Records.” *International Economic Review* 41 (1): 59–92.

Pascual-Ezama, David, Drazen Prelec, and Derek Dunfield. 2013. “Motivation, Money, Prestige and Cheats.” *Journal of Economic Behavior & Organization* 93: 367–373.

Pelham, Brett W., and Efrat Neter. 1995. “The Effect of Motivation of Judgment Depends on the Difficulty of the Judgment.” *Journal of Personality and Social Psychology* 68 (4): 581.

Phillips, James S., and Sara M. Freedman. 1988. “The Task-Related Competency and Compliance Aspects of Goal Setting: A Clarification.” *Organizational Behavior and Human Decision Processes* 41 (1): 34–49.

Phillips, James S., and Robert G. Lord. 1980. “Determinants of Intrinsic Motivation: Locus of Control and Competence Information as Components of Deci’s Cognitive Evaluation Theory.” *Journal of Applied Psychology* 65 (2): 211.

Pinder, Craig C. 1976. “Additivity versus Nonadditivity of Intrinsic and Extrinsic Incentives: Implications for Work Motivation, Performance, and Attitudes.” *Journal of Applied Psychology* 61 (6): 693.

Pokorny, Kathrin. 2008. “Pay—but Do Not Pay Too Much: An Experimental Study on the Impact of Incentives.” *Journal of Economic Behavior & Organization* 66 (2): 251–264.

Pollack, Irwin, and P. Robert Knaff. 1958. “Maintenance of Alertness by a Loud Auditory Signal.” *The Journal of the Acoustical Society of America* 30 (11): 1013–1016.

Prendergast, Canice. 1999. “The Provision of Incentives in Firms.” *Journal of Economic Literature* 37 (1): 7–63.
Pritchard, Robert D., Kathleen M. Campbell, and Donald J. Campbell. 1977. “Effects of Extrinsic Financial Rewards on Intrinsic Motivation.” *Journal of Applied Psychology* 62 (1): 9.
Pritchard, Robert D., and Michael I. Curts. 1973. “The Influence of Goal Setting and Financial Incentives on Task Performance.” *Organizational Behavior and Human Performance* 10 (2): 175–183.
Pritchard, Robert D., and Philip J. De Leo. 1973. “Experimental Test of the Valence-Instrumentality Relationship in Job Performance.” *Journal of Applied Psychology* 57 (3): 264.
Pritchard, Robert D., John Hollenback, and Philip J. DeLeo. 1980. “The Effects of Continuous and Partial Schedules of Reinforcement on Effort, Performance, and Satisfaction.” *Organizational Behavior and Human Performance* 25 (3): 336–353.
Pritchard, Robert D., Dale W. Leonard, Clarence W. Von Bergen, and Raymond J. Kirk. 1976. “The Effects of Varying Schedules of Reinforcement on Human Task Performance.” *Organizational Behavior and Human Performance* 16 (2): 205–230.

Rabin, Matthew. “Risk aversion and expected-utility theory: A calibration theorem.” *Econometrica* 68, no. 5 (2000): 1281-1292.

Rabin, Matthew, and Richard H. Thaler. “Anomalies: risk aversion.” *The Journal of Economic Perspectives* 15, no. 1 (2001): 219-232.

Raihani, N. J., Mace, R., & Lamba, S. (2013). The effect of $1, $5 and $10 stakes in an online dictator game. *PloS one*, *8*(8), e73131.

Reeson, A., & Whitten, S. (2014). *Designing Auctions for Different Environmental Commodities*. CSIRO Sustainable Agriculture Flagship.
Remus, William, Marcus O’Connor, and Kenneth Griggs. 1998. “The Impact of Incentives on the Accuracy of Subjects in Judgmental Forecasting Experiments.” *International Journal of Forecasting* 14 (4): 515–522.
Reynaud, A., & Couture, S. (2012). Stability of risk preference measures: results from a field experiment on French farmers. *Theory and decision*, *73*(2), 203-221.
Riedel, James A., Delbert M. Nebeker, and Barrie L. Cooper. 1988. “The Influence of Monetary Incentives on Goal Choice, Goal Commitment, and Task Performance.” *Organizational Behavior and Human Decision Processes* 42 (2): 155–180.

Rode, J., Gómez-Baggethun, E., & Krause, T. (2015). Motivation crowding by economic incentives in conservation policy: A review of the empirical evidence. *Ecological Economics*, *117*, 270-282.

Roth, A. E., Prasnikar, V., Okuno-Fujiwara, M., & Zamir, S. (1991). Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An experimental study. *The American Economic Review*, 1068-1095.

Rutström, E. E. (1998). Home-grown values and incentive compatible auction design. *International Journal of Game Theory*, *27*(3), 427-441.

Ryan, M., and K. Gerard 2003. “Using Discrete Choice Experiments in Health Economics: Moving Forward,” *Advances in health economics*: 25-40.

Ryan, M., and R.K. Gerard (2008). Discrete Choice Experiments in a Nutshell *Using Discrete Choice Experiments to Value Health and Health Care* (pp. 13-46): Springer.
Salvemini, Nat J., Richard R. Reilly, and James W. Smither. 1993. “The Influence of Rater Motivation on Assimilation Effects and Accuracy in Performance Ratings.” *Organizational Behavior and Human Decision Processes* 55 (1): 41–60.

Schindler, S., & Pfattheicher, S. (2017). The frame of the game: Loss-framing increases dishonest behavior. *Journal of Experimental Social Psychology*, *69*, 172-177.

Schechter, Laura. "Risk aversion and expected-utility theory: A calibration exercise." *Journal of Risk and Uncertainty* 35, no. 1 (2007): 67-76.

Schilizzi, S.G., 2017. An overview of laboratory research on conservation auctions. Land Use Policy, 63, pp.572-583.

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

Scott, W. E., Jiing-Lih Farh, and Philip M. Podsakoff. 1988. "The Effects of 'Intrinsic' and 'Extrinsic' Reinforcement Contingencies on Task Behavior." *Organizational Behavior and Human Decision Processes* 41 (3): 405–425.

Shearer, B. (2004). Piece rates, fixed wages and incentives: Evidence from a field experiment. *The Review of Economic Studies*, *71*(2), 513-534.

Shogren, J. F., Margolis, M., Koo, C., & List, J. A. (2001). A random nth-price auction. *Journal of economic behavior & organization*, *46*(4), 409-421.

Sefton, M. (1992). Incentives in simple bargaining games. *Journal of Economic Psychology*, *13*(2), 263-276.

Sipowicz, Raymond R., J. Robert Ware, and Robert A. Baker. 1962. "The Effects of Reward and Knowledge of Results on the Performance of a Simple Vigilance Task." *Journal of Experimental Psychology* 64 (1): 58.

Slonim, R., & Roth, A. E. (1998). Learning in High Stakes Ultimatum Games: An Experiment in the Slovak Republic. *Econometrica*, *66*(3), 569-596.

Slonim, R., Wang, C., Garbarino, E., & Merrett, D. (2013). Opting-in: Participation bias in economic experiments. *Journal of Economic Behavior & Organization*, *90*, 43-70.

Smith, R.D. 2003. "Construction of the Contingent Valuation Market in Health Care: A Critical Assessment," *Health economics 12*(8): 609-628.

Smith, Russell L., Luigi F. Lucaccini, and Murray H. Epstein. 1967. "Effects of Monetary Rewards and Punishments on Vigilance Performance." *Journal of Applied Psychology* 51 (5p1): 411.

Smith, Timothy W., and Thane S. Pittman. 1978. "Reward, Distraction, and the Overjustification Effect." *Journal of Personality and Social Psychology* 36 (5): 565.

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

Sostek, Andrew J. 1978. "Effects of Electrodermal Lability and Payoff Instructions on Vigilance Performance." *Psychophysiology* 15 (6): 561–568.

Steier, L. (2001). Family firms, plural forms of governance, and the evolving role of trust. *Family Business Review*, *14*(4), 353-367.

Stone, Dan N., and David A. Ziebart. 1995. "A Model of Financial Incentive Effects in Decision Making." *Organizational Behavior and Human Decision Processes* 61 (3): 250–261.

Straub, P. G., & Murnighan, J. K. (1995). An experimental investigation of ultimatum games: Information, fairness, expectations, and lowest acceptable offers. *Journal of Economic Behavior & Organization*, *27*(3), 345-364.

Straub, Tim, Henner Gimpel, and Florian Teschner. 2014. "The Negative Effect of Feedback on Performance in Crowd Labor Tournaments," 4.

Sydnor, Justin. "(Over) insuring modest risks." *American Economic Journal: Applied Economics* 2, no. 4 (2010): 177-199.

Takahashi, Hiromasa, Junyi Shen, and Kazuhito Ogawa. 2016. "An Experimental Examination of Compensation Schemes and Level of Effort in Differentiated Tasks." *Journal of Behavioral and Experimental Economics* 61: 12–19.

Tanaka, T., Camerer, C. F., and Nguyen, Q. (2010). Risk and time preferences: linking experimental and household survey data from Vietnam. *American Economic Review*, *100*(1), 557–571.

Terborg, James R., and Howard E. Miller. 1978. "Motivation, Behavior, and Performance: A Closer Examination of Goal Setting and Monetary Incentives." *Journal of Applied Psychology* 63 (1): 29.

Thaler, R. H. (1988). Anomalies: The winner's curse. *Journal of Economic Perspectives*, *2*(1), 191-202.

Tisdell, J.G. and Iftekhar, M.S., 2013. Fisheries quota allocation: Laboratory experiments on simultaneous and combinatorial auctions. Marine Policy, 38, pp.228-234.

Toler, S., Briggeman, B. C., Lusk, J. L., & Adams, D. C. (2009). Fairness, farmers markets, and local production. *American Journal of Agricultural Economics*, *91*(5), 1272-1278.

Tompkinson, P., & Bethwaite, J. (1995). The ultimatum game: raising the stakes. *Journal of Economic Behavior & Organization*, *27*(3), 439-451.

Tomporowski, Phillip D., Royce G. Simpson, and Lisa Hager. 1993. "Method of Recruiting Subjects and Performance on Cognitive Tests." *The American Journal of Psychology*, 499–521.

Tonin, Mirco, and Michael Vlassopoulos. 2013. "Social Incentives Matter: Evidence from an Online Real Effort Experiment."

Van Dijk, Frans, Joep Sonnemans, and Frans Van Winden. 2001. "Incentive Systems in a Real Effort Experiment." *European Economic Review* 45 (2): 187–214.

Vandegrift, Donald, Abdullah Yavas, and Paul M. Brown. 2007. "Incentive Effects and Overcrowding in Tournaments: An Experimental Analysis." *Experimental Economics* 10 (4): 345–368.

Vecchio, Robert P. 1982. "The Contingent-Noncontingent Compensation Controversy: An Attempt at a Resolution." *Human Relations* 35 (6): 449–462.

Velez, M. A., Stranlund, J. K., & Murphy, J. J. (2009). What motivates common pool resource users? Experimental evidence from the field. *Journal of Economic Behavior & Organization*, *70*(3), 485-497.

Venkatachalam, L. (2008). Behavioral economics for environmental policy. *Ecological Economics*, *67*(4), 640-645.

Vickrey, W. (1961). Counterspeculation, auctions, and competitive sealed tenders. *The Journal of finance*, *16*(1), 8-37.

Viesti, Carl R. 1971. "Effect of Monetary Rewards on an Insight Learning Task." *Psychonomic Science* 23 (2): 181–183.

Vogt, N., Reeson, A.F. and Bizer, K., 2013. Communication, competition and social gift exchange in an auction for public good provision. Ecological economics, 93, pp.11-19.

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.

Waller, William S., and Chee W. Chow. 1985. "The Self-Selection and Effort Effects of Standard-Based Employment Contracts: A Framework and Some Empirical Evidence." *Accounting Review*, 458–476.

Wasserman, Edward A., Bernard Weiner, and John P. Houston. 1968. "Another Failure for Motivation to Enhance Trace Retrieval." *Psychological Reports* 22 (3): 1007–1008.

Weiner, Bernard. 1966. “Motivation and Memory.” *Psychological Monographs: General and Applied* 80 (18): 1.
Weiner, Michael Jay, and Anthony M. Mander. 1978. “The Effects of Reward and Perception of Competency upon Intrinsic Motivation.” *Motivation and Emotion* 2 (1): 67–73.
Wickens, Delos D., and C. Kenneth Simpson. 1968. “Trace Cue Position, Motivation, and Short-Term Memory.” *Journal of Experimental Psychology* 76 (2p1): 282.
Wiener, Earl L. 1969. “Money and the Monitor.” *Perceptual and Motor Skills* 29 (2): 627–634.
Wimperis, Bruce R., and James L. Farr. 1979. “The Effects of Task Content and Reward Contingency upon Task Performance and Satisfaction.” *Journal of Applied Social Psychology* 9 (3): 229–249.
Wright, Patrick M. 1989. “Test of the Mediating Role of Goals in the Incentive-Performance Relationship.” *Journal of Applied Psychology* 74 (5): 699.
Wright, P. M.. 1990. “Monetary Incentives and Task Experience as Determinants of Spontaneous Goal Setting, Strategy Development, and Performance.” *Human Performance* 3 (4): 237–258.
Wright, Patrick M., and K. Michele Kacmar. 1995. “Mediating Roles of Self-Set Goals, Goal Commitment, Self-Efficacy, and Attractiveness in the Incentive-Performance Relation.” *Human Performance* 8 (4): 263–296.
Wright, William F., and Mohamed E. Aboul-Ezz. 1988. “Effects of Extrinsic Incentives on the Quality of Frequency Assessments.” *Organizational Behavior and Human Decision Processes* 41 (2): 143–152.
Wright, William F., and Urton Anderson. 1989. “Effects of Situation Familiarity and Financial Incentives on Use of the Anchoring and Adjustment Heuristic for Probability Assessment.” *Organizational Behavior and Human Decision Processes* 44 (1): 68–82.

Yue, C., and C. Tong 2009. “Organic or Local? Investigating Consumer Preference for Fresh Produce Using a Choice Experiment with Real Economic Incentives,” *HortScience 44*(2): 366-371.
Yukl, Gary, Kenneth N. Wexley, and James D. Seymore. 1972. “Effectiveness of Pay Incentives under Variable Ratio and Continuous Reinforcement Schedules.” *Journal of Applied Psychology* 56 (1): 19.
Zivin, Joshua S. Graff, Lisa B. Kahn, and Matthew J. Neidell. 2019. “Incentivizing Learning-By-Doing: The Role of Compensation Schemes.” National Bureau of Economic Research.

Appendices

A.1 List of Studies Using Real Effort Tasks

Table A1. Real effort task studies reviewed

Type/author	Subs	Txs	Subs/Tx	Incentive Type	Task	Payment	Incentive Effect	Issues
Vigilance Tasks								
(Bahrick, Fitts, and Rankin 1952)	100	9	50	F/Q	Track	F: $0.75/hr, Q: F+ $0.05-$3.00 for good perf.	+	2
(Bergum and Lehr 1964)	40	2	20	N/P	Detect	P: $0.2 correct, -$0.2 incorrect	+/=	1
(Bevan and Turner 1965)	40	2	20	N/P	Detect	P: $0.01 correct	+	2
(Glucksberg 1962)	128	4	32	N/T	Read	T: $20 best, $5 top 25%	-/+	2
(Montague and Webber 1965)	25	2	15,10	N/P	Detect	P: $0.01 correct, -$0.01 incorrect	+	1
(Pollack and Knaff 1958)	10	2	5	F/Q	Detect	Q: Unkonwn	=	2
(Sipowicz, Ware, and Baker 1962)	37	2	18,19	N/Q	Detect	Q $3.00 if 100%, deduct for miss starting at $0.05	+	1
(R. L. Smith, Lucaccini, and Epstein 1967)	48	6	8	N/P	Detect	P: $0.1/$0.2 correct, -$0.1/$0.2 incorrect,	=/+	1
(Sostek 1978)	66	3	22	N/P	Detect	P: $0.09 correct, -$0.09/$0.01 miss, $-0.01/$0.09 false alarm	+	1
(Tomporowski, Simpson, and Hager 1993) (exp I)	70	3	16-21	N/F	Detect	F: $5/$10 hour	=/+	1
(Tomporowski, Simpson, and Hager 1993) (exp II)	40	2	20	N/F	Match	F: $10 hour	+	1
(Wiener 1969)	15	2	7,8	N/P	Detect	P: $0.05 correct, -$0.05 incorrect	=	1
Memory								
(Bahrick 1954)	74	2	37	N/Q	Learn	Q: $0.1 - $1.5, $0.1 incorrect	+	2
(Dornbush 1965) (exp I)	60	3	20	N/P	Learn	Unknown	+	3
(Dornbush 1965) (exp II)	60	3	20	N/P	Learn	Unknown	=	3
(W. F. Harley 1968)	180	9	20	N/P	Recall	P: $0.05 correct	=/+	1
(W. Harley 1965)(exp a)	40	4	10	N/P	Recall	P: $0.25 correct	=/+	1
(W. Harley 1965) (exp	80	4	20	N/P	Recall	P: $0.25 correct	=	1

Type/author	Subs	Txs	Subs/Tx	Incentive Type	Task	Payment	Incentive Effect	Issues
Vigilance Tasks								
b)								
(Hell et al. 1988)	60	2	28-31	T/F	Recall	T: 25 deutsch marks for top three	=/+	1
(Kausler and Trapp 1962)	80	8	10	N/Q	Learn	Q: $0.50 - $2.50 correct	=/+	1
(Kernoff, Weiner, and Morrison 1966)	20	7	3	N/P	Recall	P: $0.01/$0.05 correct	=	1
(Libby and Lipe 1992)	117	3	38-40	F/P, F/T	Recall	F: $2 participation, P: F + $0.1/correct, T: F + $0.1/correct + top 5 $5.00 bonus	+	
(McNamara and Fisch 1964)	20	5	4	F/N	Various	Differing rates per task	-/+	3
(Nilsson 1987) (exp 1)	30	3	10	N/T	Recall	T: $10 for best	=	1
(Nilsson 1987) (exp 2)	30	3	10	N/T	Recall	T: $10 for best	=	1
(Pritchard, Hollenback, and DeLeo 1980)	60	3	20	F/P, F/V	Learn	F: $2.00/hour, P: depending on book, average $2.00/hour, V: $2.00/hour times $0 - $6	+	1
(Pritchard et al. 1976)	24	2	8-16	F/P, F/V	Learn	F: $2/hr. P (between sub design) P: $3/3 tests, V: $3.00 on average	+	1
(Tomporowski, Simpson, and Hager 1993) (exp 3)	60	3	20	F/N	Recall words	F: $5/$10 hour	=	1
(Wasserman, Weiner, and Houston 1968)	32	4	8	N/P	Recall	P: $0.05 correct	=	3
Weiner (1966) (exp 3)	20	4	5	N/P	Recall	P: $0.05 correct	+	1
Weiner (1996) (exp 10)				N/P	Recall	P:$0.01/$0.05 correct	=/+	1
Weiner (1996) (exp 11)	57	6	18-21	N/P	Recall	P: $0.05 correct	=/+	1
Weiner (1996) (exp 12)	57	6	18-21	N/P	Recall	P: $0.05 correct	=/+	1
Weiner (1996) (exp 13)	72	4	18	N/P	Recall	P: $0.05 correct	=/+	1
Weiner (1996) (exp 14)	72	4	18	N/P	Recall	P: $0.05 correct	=/+	1
Weiner (1996) (exp 15)	72	4	18	N/P	Recall	P: $0.05 correct	=/+	1
Weiner (1996) (exp 2)	20	4	5	N/P	Recall	P: $0.05 correct	=	1
Weiner (1996) (exp 4)	20	4	5	N/P	Recall	P: $0.05 correct	=	1
Weiner (1996) (exp 6)	16	4	4	N/P	Recall	P:$0.01/$0.05 correct	=	1
Weiner (1996) (exp 7)	16	2	8	N/P	Recall	P: $0.05 correct	=	1
Weiner (1996) (exp 8)	16	2	8	N/P	Recall	P: $0.05 correct	=	1

Type/author	Subs	Txs	Subs/Tx	Incentive Type	Task	Payment	Incentive Effect	Issues
Vigilance Tasks								
Weiner (1996) (exp 9)	164	6	27	N/P	Recall	P: $0.05 correct	=	1
(Wickens and Simpson 1968)	192	4	48	N/P	Recall	P: $0.05 correct	=/+	1
Production and Clerical Tasks								
(Adams and Rosenbaum 1962) (exp 2)	58	4	18-Nov	F/P	Interview	F: $3.50/hour, P: 30c/piece	-/=	1
(Al-Ubaydli et al. 2015)	78	2	47,31	F/P	Stuff	F: $9 hr, P: $8 hr+ $0.2/envelope	=/+	2
(Atkinson and Reitman 1956)	93	2	49, 43	N/T	Draw	T: $5/highest score	+	2
(Bailey, Brown, and Cocco 1998)	72	3	24	F/P, F/Q	Assemble	F: $20, P: $1.80/unit, Q: flat base of $17.50, plus bonus of $3 or $6	+	2
(Byram 1997) (exp 5)	64	2	34, 32	T/F	Fold	F: $3 T: $4 1st quartile, $2 top half, $1 top 75%, $0 otherwise	=	2
(Carpenter, Matthews, and Schirm 2010)	106	2	53	P/T	Stuff	P: $1 quality envelope, T: P + $25 winner	+	
(Charness, Cobo-Reyes, and Sanchez 2016)	90	3	30	N/P/P	Data entering	P: $0/$0.02/$0.08	+	1
(Chung and Vickery 1976)	80	8	10	F/P	Transfer	F: $2 hr + information, P: $0.35c page + $0.05 column last page + -$0.05 error + chance to win $2 or $4	=/+	2
(Dillard and Fisher 1990)	27	2	13-14	F/Q	Decode	F: $3, Q: $2 + $0.54*(output over standard)	+	
(Erez, Gopher, and Arzi 1990)	16	2	8	N/Q	Type	Q: $0.25 for 0.1 better standard score (max $1.5)	=/+	
(Farh, Griffeth, and Balkin 1991)	65	2	27	F/Q	Decode	F: $3.5 half hr, Q: $1 fail or $5 success, P: $0.05/$0.10/$0.15 per line	+	1
(Farh, Griffeth, and Balkin 1991)	54	6	9	P/Q	Decode	F: $3.50, P: 5c/10c/15c per card, Q: $1 if below norm, $5 bonus for exceeding	+	1
(Farr 1976)	90	6	15	F/P	Assemble	P: $0.15/$0.35 unit	+	1
(Fatseas and Hirst 1992)	180	12	15	F/P	Decode	F: $5, P: $0.28 line (approx $5), Q: $1 participation, bonus for	+/=	

Type/author	Subs	Txs	Subs/Tx	Incentive Type	Task	Payment	Incentive Effect	Issues
Vigilance Tasks								
						success		
(Fest et al. 2019)	897	3	295-302	P/P	Transcript	P: $0/$0.01/$0.05	+/=	1
(Frisch and Dickinson 1990)	75	5	15	F/Q	Assemble	Q: $4 + 0/10/30/60/100% above base pay for success.	+	2
(Hamner and Foster 1975)	98	6	15-20	N/F/P	Transfer	F: $0.75, P: $0.05 scored survey	=/+	4
(Henry and Sniezek 1993)	240	12	20	N/T	Answer	T: $50, $40, $30, $20, $10 for top 5 finishers	=	1
(Henry and Strickland 1994)	137	4	34	N/T	Answer	T: $25 1st place, $20 for 2nd-5th, $10 for 6th-10th for entire study (many sessions)	+	1
(Huber 1985)	88	6	14 - 15	F/Q	Proof read	Several Q schemes	=/+	3
(Jorgenson and Dunnette 1973)	256	6	18-48	F/Q	Find	F: $2.00 hour, Q: $1.60/$2.00/$2.40	+	2
(Locke, Bryan, and Kendall 1968) (exp 2)	30	2	15	F/P	Assemble	P: $0.12 toy, F: $3	=	1
(London and Oldham 1976)	120	12	10	F/P	Sort	F: $2, P: $0.01 card , P: $0.01 card	=	1
(London and Oldham 1977)	70	5	14	F/P	Sort	F: $0.5, P: $0.01 card	+	1
(Phillips and Freedman 1988)	102	6	17	F/Q	Proof read	F: $5 participation, Q: $2 + $3 success	+	2
(Pinder 1976)	80	4	20	F/P	Construct, assemble	F: $2.75, P: $0.05 piece	=/+	
(Pritchard and Curts 1973)	81	5	10-21	N/Q	Sort	Q: $0.5/$3 success, F: $0.03/card	=/+	1
(Pritchard and De Leo 1973)	60	4	15	F/P	Decode	F: $1.75/$2.5, P: $0.07/$0.10 piece	=/+	1
(Riedel, Nebeker, and Cooper 1988)	130	7	18	F/Q	Transfer	F: $4.4 hour, Q: 25/50/75/100/125% of wage	+	1
(Scott, Farh, and Podsakoff 1988)	96	8	12	N/P	Assemble	P: 8c/piece for complex, 6c/piece for easy	+	1
(Terborg and Miller 1978)	60	6	10	F/P	Assemble	F: $2.50, P: $0.4 model	+	
(Vecchio 1982)	43	2	20, 23	F/P	Survey	F: $4.80/hr, P: $0.4/survey	+	1
(Waller and Chow 1985)	61	11	1-18	F/Q	Decode	F: $3, Q: $0.18 - $1.8	=	1
(M. J. Weiner and Mander 1978)	90	9	10	F/N/P	Decode	F: not disclosed, P: $0.05 word	=	1
(Wimperis and Farr	48	6	8	F/N/P	Assemble	N: Course credit, F:	=/+	1

Type/author	Subs	Txs	Subs/Tx	Incentive Type	Task	Payment	Incentive Effect	Issues
Vigilance Tasks								
1979)					, build	$1.75, P: Unkonwn		
(P. M. Wright 1989)	243	9	27	F/Q	Sort	F: "yoked" to sub in P, P: $0.04 card, Q: $5 + min bonus of $5.05 success	=	1
(Yukl, Wexley, and Seymore 1972)	15	3	5	V/P	Grade	F: $1.50/h, P: F + 25c/sheet, Variable-Rate: F + 50-50 chance 25c/sheet, Variable-Rate: F + 50-50 chance $0.5/sheet	+/=	4
Judgement and Choice								
(Arkes, Dawes, and Christensen 1986) (exp 1)	226	12	18	N/P/T	Judge	P: 10c correct answer, T: $5 best	-	1
(Ashton 1990)	182	8	22	N/T	Predict	T: $100 best	=/+	1
(Awasthi and Pratt 1990)	70	2	35	F/P	Decide	F: $1, P: $1 + $2 correct answer	=	1
(El-Gamal and Grether 1995)	257	2	128	F/P	Judge	F: unkonwn, P: $10 correct answer	+	1
(Gigerenzer, Hoffrage, and Kleinbölting 1991)	80	8	10	T/F	Answer	T: 20 German marks best	=/-	1
(Grether 1980)	341	6	56	F/Q	Judge	F: $5, Q: $15 success	=	1
Hogarth et al. (1991) (exp 1)	121	6	20	N/Q	Predict	Q: 1c for points above a threshold	=/+	2
Hogarth et al. (1991) (exp 3)	80	4	20	N/P	Predict	P: non-linear/ unkonwn	=	1
Hogarth et al. (1991) (exp 5)	90	4	22	N/Q	Predict	Q: $0.01 above a threshold	=	2
(Pelham and Neter 1995) (exp 3)	85	4	21	T/N	Estimate	T: $25 to 10 most accurate	=/+	1
(Remus, O'Connor, and Griggs 1998)	51	4	12	N/F/T	Forecast	F: $5, T: $20 for 1st, $15 2nd, $10 3rd	=	2
(Salvemini, Reilly, and Smither 1993)	180	9	20	T/N	Judge	T: $200, $150, $75, $50, $25 for 1st-5th	+	1
(Stone and Ziebart 1995)	84	2	42	F/Q	Choose	F: random distribution of payment, Q: $10 100%	+	1
Viesti (1971) (exp 1)	32	2	16	N/P	Judge	P: $0.50 judgement	-	1
Viesti (1971) (exp 2)	61	2	30.5	N/P	Judge	P: $1 judgment	=	1
(W. F. Wright and Aboul-Ezz 1988)	51	4	29, 22	F/T	Predict	F: $3, T: $25 1st, $20 for 2nd-3rd, $15 for 4th-5th	+	1

Type/author	Subs	Txs	Subs/Tx	Incentive Type	Task	Payment	Incentive Effect	Issues
Vigilance Tasks								
(W. F. Wright and Anderson 1989) (exp 3)	160	6	26	N/T	Judge	T: $15 1st-5th, $12 6th-10th, $8 11th-15th, $6 16th-25th, $4 26th-35th, $2 36th-45th	=	2
Problem Solving, Reasoning, and Game Playing								
(Allan, Bender, and Theodossiou 2017)	40	2	20	F/P	Multiply	F: $6.34, P: $0.25/correct answer	=	
(Araujo et al. 2016)	148	3	42,43,63	P/P	Slider task	P:$0.05, $0.2, $0.8	+	1
(ariely et al. 2009)	87	3	29	Q/Q	Pack, memory, motor skills	Q: 4, 40, or 400 Indian Rs. "Very good" 100% Q, "good" 50%, or 0% below "good". Rs 400 close to av. monthly per capita consumer expenditure in rural areas.	=/-	1
(Baumeister 1984) (exp 5)	37	4	9	N/Q	Play	Q: $1 success	=	3
(Bellemare, Lepage, and Shearer 2010)	82	2	40-42	F/P	Data entering	$10 show up fee, F: $10, P: $0.10/entry	N/A	
(Bracha, Gneezy, and Loewenstein 2015)	88	2	44	P/P	Solve	P: $0.40/$0.80 correct answer	+	
(Cadsby, Song, and Tapon 2007)	115	2	57-58	F/P	Find	F: $1.57/round, P: $0.14/word	+	
(Campbell 1984)	56	6	9	F/Q	Solve	F: $3, Q: $0.60 success (up to $3)	+	
(Carpenter and Gong 2016)	207	6	34-35	F/P, P/P	Stuffing	F: $20/15 minutes, P: F + $0.5 letter/ F + $1 letter	+	
(Cason, Masters, and Sheremeta 2010)	138	2	69	P/T	Add	P: $0.40/answer, T: $20 winner	N/A	
(Corgnet, Gómez-Miñambres, and Hernán-Gonzalez 2015)	94	3	31	P/P	Sum	P: $0.10, $0.8, $1.50	+/=	1
(Dalton, Gonzalez Jimenez, and Noussair 2016)	232	3	39,93	P/P	Count	P: E0.20/0.50 correct, -E0.20/-0.50 third incorrect	=	2
(DellaVigna and Pope 2017)	2226	4	540-556	P/P	Press buttons	P: $0/$0.01/$0.04/$0.1	+	
(Dohmen and Falk 2011)	360	3	120	F/P/T	Multiply	F: $52, P:$0.22/correct answer or $0.22 shared by team, T: $0 vs $29	+	self-sort into incentive

Type/author	Subs	Txs	Subs/Tx	Incentive Type	Task	Payment	Incentive Effect	Issues
Vigilance Tasks								
								scheme
(Eckartz, Kirchkamp, and Schunk 2012)	216	3	216	F/P/T	Solve, answer, add	F: $11, P: increasing rate per length of words, T: $28 for winning, 6 for losing	=	1
(Enzle and Ross 1978)	72	6	12	N/Q	Solve	Q: $0.45/$1.50 success	=	1
(Eriksson, Poulsen, and Villeval 2008)	208	6	28-48	P/T	Add	P: $0.15/answer , T: $25 winner	N/A	
Erkal, Gangadharan, and Koh (2018) Exp 1	156	3	52	F/T, T/T	Encrypt	$10 show-up, F: 15 AUD, T: 25/60 AUD winner, 15 loser	+/=	
Erkal, Gangadharan, and Koh (2018) Exp 2	161	3	54	F/T, T/T	Encrypt	$10 show-up, F: 15 AUD, T: 25/60 AUD winner, 15 loser	+	
Farr, Vance, and McIntyre (1977) (exp 1)	48	6	8	F/P	Solve	F: $1/$2/$3, P: $0.5/$1/$1.50 puzzle	=	1
Farr, Vance, and McIntyre (1977) (exp 2)	152	6	25	F/P	Solve	F: $2, P: $0.70 puzzle	=	1
(Fehrenbacher and Pedell 2012)	165	6	21-38	F/P/Q	Solve	F: $12.8, P: $0.3/anagram, Q: $35.8 target, $5.1 otherwise	N/A	
(Freeman and Gelber 2010)	468	6	78	F/P/T/T	Solve	Show-up fee: $13, F: $5 to all, P: $0.20/maze, T: $30 only winner, $15 /$7/$5/$2/$1	+/=	
(Friedl, Neyse, and Schmidt 2018)	114	2	57	F/P	Slider task	F: 3.75 pounds, P: 0.03 pounds per slider	+	
(Gächter, Huang, and Sefton 2016) study 1	20	2	10	P/P	Click	$4.6 show up fee, P: $0.01/$0.03	+	
(U. Gneezy, Niederle, and Rustichini 2003)	216	2	108	P/T	Solve	Show up $4.88, P: $0.5/maze, T: $3 winner	N/A	1
(Uri Gneezy and Rustichini 2000) exp 1	160	4	40	N/P, P/P	Answer	P: $0.03/$0.3/$0.86 correct answer	+/=	
(Goerg, Kube, and Radbruch 2017)	48	2	24	P/P	Slider task	P: $0.022/0.11 per screen	=	
(Greiner, Ockenfels, and Werner 2011)	30	2	15	P/P	Data entering	P: $0.1, $0.25, $0.4	+	
(Hennig-Schmidt, Sadrieh, and Rockenbach 2010)	20	2	10	P/P	Stuffing	show up fee: $2, P: $3.3/$3.7	=	1
(Johnson and Dickinson 2010)	26	2	6-20	F/T	Data entering	F: $5.25, T: F + $50 winner	N/A	4

Type/author	Subs	Txs	Subs/Tx	Incentive Type	Task	Payment	Incentive Effect	Issues
Vigilance Tasks								
(Kachelmeier, Reichert, and Williamson 2008)	40	2	20	F/P	Solve	F: $25, P: unknown	+	1
(Lee, Locke, and Phan 1997)	102	6	17	P/F, T/F	Solve	F: $9, P: $0.1 problem ($3 minimum), Q: $3 + $1/$3/$5 depending on goal	=	1
(Locke, Bryan, and Kendall 1968)l (exp 1)	127	7	18	F/P/Q	List	F: $3, P: $0.006/$0.004, Q: 10c/25c success	=	1
(McGraw and McCullers 1979)	72	2	36	N/Q	Solve	Q: $0.5 answer + $1 100%	=/-	1
(Pascual-Ezama, Prelec, and Dunfield 2013)	120	6	20	T	Find	T: best performers receive $11.19 extra	+	1
(Phillips and Lord 1980)	56	4	14	F/Q	Play	F: $2, Q: $1 success	=	2
(Pokorny 2008)	132	4	33	N/P	Count	$5 show-up fee, P: E0.01/E0.05/E0.5 point scored	+/-	
(Pritchard, Campbell, and Campbell 1977)	28	4	7	N/T	Solve	T: Best performer gets $5	=	
(Rubin, Samek, and Sheremeta 2016) part 1	197	3	42-100	F/P, P/P	Add	P: $0.05, $0.25, $1, $3 per correct	+/-	
(T. W. Smith and Pittman 1978)	132	4	33	N/Q	Solve	Q: Varying	=	2
(Straub, Gimpel, and Teschner 2014)	331	4	74-97	P/T	Slider task	Show up $0.30, P: $0.01/slider, T: $1 winner	N/A	1
(Takahashi, Shen, and Ogawa 2016)	145	5	24-34	N, F/F, P/P	Click, solve	N: no pay, P: $0.005, $0.02 for circles, $1/$4 for puzzles.	=/+, +	
(Tonin and Vlassopoulos 2013)	104	2	52	P/P	Data entering	P: $0.03, $0.03/$0.06/$0.10/$0.13	?	1
(Van Dijk, Sonnemans, and Van Winden 2001)	79	3	24-28	P/T	Find	P: variable, T: $0.6 win, $0.15 lose, $0.37 tie	N/A	2
(Vandegrift, Yavas, and Brown 2007)	180	4	45	P/T	Forecast	P: variable, T: $4.5 winner, $2.25/$1.5/$0.75/$0	N/A	
(P. M. Wright 1990)	245	8	30.625	N/F/P/T	Develop	F: yoked amount to a P sub, P: $0.75 schedule, T: $3.75 + bonus between $3.75-6.75 if in top 1/3	+	2

Type/author	Subs	Txs	Subs/Tx	Incentive Type	Task	Payment	Incentive Effect	Issues
Vigilance Tasks								
(P. M. Wright and Kacmar 1995) (exp 1)	80	5	16	P/F, T/F	Solve	P: $0.75 anagram, T: $3 top 3	=/+	1

Notes: No monetary rewards (N), Fixed (F), Piece-rates (P), Tournaments (T), and Variable-rate (V) payments. Incentive type columns shows the study's compared incentive types. Issues: missing relevant statistics (1), confounding factors (2), cannot source article (3), deception (4).

A.2 List of Studies using Holt and Laury Risk Elicitation Tasks

1. Anderson, L. R., & Mellor, J. M. (2008). Predicting health behaviors with an experimental measure of risk preference. *Journal of health economics*, *27*(5), 1260-1274.
2. Anderson, L. R., & Mellor, J. M. (2009). Are risk preferences stable? Comparing an experimental measure with a validated survey-based measure. *Journal of Risk and Uncertainty*, *39*(2), 137-160.
3. Bellemare, C., & Shearer, B. (2010). Sorting, incentives and risk preferences: Evidence from a field experiment. *Economics Letters*, *108*(3), 345-348.
4. Bosch-Domènech, A., & Silvestre, J. (2013). Measuring risk aversion with lists: a new bias. *Theory and decision*, *75*(4), 465-496.
5. Brown, A. L., & Kim, H. (2013). Do individuals have preferences used in macro-finance models? An experimental investigation. *Management Science*, *60*(4), 939-958.
6. Carlsson, F., Martinsson, P., Qin, P., & Sutter, M. (2013). The influence of spouses on household decision making under risk: an experiment in rural China. *Experimental Economics*, *16*(3), 383-401.
7. Charness, G., & Viceisza, A. (2012). Comprehension and risk elicitation in the field: Evidence from rural Senegal.
8. Chen, Y., Katuščák, P., & Ozdenoren, E. (2013). Why can't a woman bid more like a man?. *Games and Economic Behavior*, *77*(1), 181-213.
9. Crosetto, P., & Filippin, A. (2016). A theoretical and experimental appraisal of four risk elicitation methods. *Experimental Economics*, *19*(3), 613-641.
10. Dave, C., Eckel, C. C., Johnson, C. A., & Rojas, C. (2010). Eliciting risk preferences: When is simple better?. *Journal of Risk and Uncertainty*, *41*(3), 219-243.
11. Deck, C., Lee, J., Reyes, J. A., & Rosen, C. C. (2013). A failed attempt to explain within subject variation in risk taking behavior using domain specific risk attitudes. *Journal of Economic Behavior & Organization*, *87*, 1-24.
12. Deck, C. A., Lee, J., Reyes, J. A., & Rosen, C. (2008). Measuring risk attitudes controlling for personality traits. *Available at SSRN 1148521*.
13. Galbiati, R., & Vertova, P. (2008). Obligations and cooperative behaviour in public good games. *Games and Economic Behavior*, *64*(1), 146-170.

14. Goeree, J. K., Holt, C. A., & Palfrey, T. R. (2003). Risk averse behavior in generalized matching pennies games. *Games and Economic Behavior*, *45*(1), 97-113.
15. Grijalva, T., Berrens, R. P., & Shaw, W. D. (2011). Species preservation versus development: An experimental investigation under uncertainty. *Ecological Economics*, *70*(5), 995-1005.
16. Harrison, G. W., Lau, M. I., Rutström, E. E., & Tarazona-Gómez, M. (2012). Preferences over social risk. *Oxford Economic Papers*, *65*(1), 25-46.
17. Herberich, D. H., & List, J. A. (2012). Digging into background risk: experiments with farmers and students. *American Journal of Agricultural Economics*, *94*(2), 457-463.
18. Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *American economic review*, *92*(5), 1644-1655.
19. Jacquemet, N., Rullière, J. L., & Vialle, I. (2008). Monitoring optimistic agents. *Journal of Economic Psychology*, *29*(5), 698-714.
20. Laury, S. (2005). Pay one or pay all: Random selection of one choice for payment. *Andrew Young School of Policy Studies Research Paper Series*, (06-13).
21. Loomes, G., & Pogrebna, G. (2014). Measuring individual risk attitudes when preferences are imprecise. *The Economic Journal*, *124*(576), 569-593.
22. Lusk, J. L., & Coble, K. H. (2005). Risk perceptions, risk preference, and acceptance of risky food. *American Journal of Agricultural Economics*, *87*(2), 393-405.
23. Nielsen, T., Keil, A., & Zeller, M. (2013). Assessing farmers' risk preferences and their determinants in a marginal upland area of Vietnam: a comparison of multiple elicitation techniques. *Agricultural Economics*, *44*(3), 255-273.
24. Reynaud, A., & Couture, S. (2012). Stability of risk preference measures: results from a field experiment on French farmers. *Theory and decision*, *73*(2), 203-221.
25. Schunk, D. (2009). Behavioral heterogeneity in dynamic search situations: Theory and experimental evidence. *Journal of Economic Dynamics and Control*, *33*(9), 1719-1738.

A.3 List of Studies using Eckel Grossman Risk Elicitation Tasks

1. Ball, S., Eckel, C. C., & Heracleous, M. (2010). Risk aversion and physical prowess: Prediction, choice and bias. *Journal of Risk and Uncertainty*, *41*(3), 167-193.
2. Dave, C., Eckel, C. C., Johnson, C. A., & Rojas, C. (2010). Eliciting risk preferences: When is simple better?. *Journal of Risk and Uncertainty*, *41*(3), 219-243.
3. Deck, C., Lee, J., Reyes, J. A., & Rosen, C. C. (2013). A failed attempt to explain within subject variation in risk taking behavior using domain specific risk attitudes. *Journal of Economic Behavior & Organization*, *87*, 1-24.
4. Eckel, C. C., El-Gamal, M. A., & Wilson, R. K. (2009). Risk loving after the storm: A Bayesian-Network study of Hurricane Katrina evacuees. *Journal of Economic Behavior & Organization*, *69*(2), 110-124.
5. Eckel, C. C., & Grossman, P. J. (2002). Sex differences and statistical stereotyping in attitudes toward financial risk. *Evolution and human behavior*, *23*(4), 281-295.
6. Eckel, C. C., & Grossman, P. J. (2008). Forecasting risk attitudes: An experimental study using actual and forecast gamble choices. *Journal of Economic Behavior & Organization*, *68*(1), 1-17.

7. Engle-Warnick, J., Escobal, J., & Laszlo, S. (2009). How do additional alternatives affect individual choice under uncertainty?. *Canadian Journal of Economics/Revue canadienne d'économique*, *42*(1), 113-140.

A.4 List of Studies using Certain vs. Risky Elicitation Tasks

1. Balafoutas, L., Kerschbamer, R., & Sutter, M. (2012). Distributional preferences and competitive behavior. Journal of economic behavior & organization, 83(1), 125-135.
2. Barham, B. L., Chavas, J. P., Fitz, D., Salas, V. R., & Schechter, L. (2014). The roles of risk and ambiguity in technology adoption. Journal of Economic Behavior & Organization, 97, 204-218.
3. Bruner, D. M. (2009). Changing the probability versus changing the reward. Experimental Economics, 12(4), 367-385.
4. Cesarini, D., Dawes, C. T., Johannesson, M., Lichtenstein, P., & Wallace, B. (2009). Genetic variation in preferences for giving and risk taking. The Quarterly Journal of Economics, 124(2), 809-842.
5. Csermely, T., & Rabas, A. (2016). How to reveal people's preferences: Comparing time consistency and predictive power of multiple price list risk elicitation methods. Journal of risk and uncertainty, 53(2-3), 107-136.
6. Fellner, G., & Maciejovsky, B. (2007). Risk attitude and market behavior: Evidence from experimental asset markets. Journal of Economic Psychology, 28(3), 338-350.
7. Gillen, B., Snowberg, E., & Yariv, L. (2015). Experimenting with measurement error: Techniques with applications to the caltech cohort study (No. w21517). National Bureau of Economic Research.
8. Gonzalez, R., & Wu, G. (1999). On the shape of the probability weighting function. Cognitive psychology, 38(1), 129-166.
9. Heinemann, F., Nagel, R., & Ockenfels, P. (2009). Measuring strategic uncertainty in coordination games. The Review of Economic Studies, 76(1), 181-221.
10. Kamas, L., & Preston, A. (2012). The importance of being confident; gender, career choice, and willingness to compete. Journal of Economic Behavior & Organization, 83(1), 82-97.
11. Kleinlercher, D., Huber, J., & Kirchler, M. (2014). The impact of different incentive schemes on asset prices. European Economic Review, 68, 137-150.
12. Koudstaal, M., Sloof, R., & Van Praag, M. (2015). Risk, uncertainty, and entrepreneurship: Evidence from a lab-in-the-field experiment. Management Science, 62(10), 2897-2915.
13. Nosić, A., & Weber, M. (2010). How riskily do I invest? The role of risk attitudes, risk perceptions, and overconfidence. Decision Analysis, 7(3), 282-301.
14. Noussair, C. N., Trautmann, S. T., & Van de Kuilen, G. (2013). Higher order risk attitudes, demographics, and financial decisions. *Review of Economic Studies*, *81*(1), 325-355.
15. Rudorf, S., Preuschoff, K., & Weber, B. (2012). Neural correlates of anticipation risk reflect risk preferences. Journal of Neuroscience, 32(47), 16683-16692.
16. Shupp, R., Sheremeta, R. M., Schmidt, D., & Walker, J. (2013). Resource allocation contests: Experimental evidence. Journal of Economic Psychology, 39, 257-267.
17. Sokol-Hessner, P., Hsu, M., Curley, N. G., Delgado, M. R., Camerer, C. F., & Phelps, E. A. (2009). Thinking like a trader selectively reduces individuals' loss aversion. Proceedings of the National Academy of Sciences, 106(13), 5035-5040.

18. Tymula, A., Belmaker, L. A. R., Roy, A. K., Ruderman, L., Manson, K., Glimcher, P. W., & Levy, I. (2012). Adolescents' risk-taking behavior is driven by tolerance to ambiguity. Proceedings of the National Academy of Sciences, 109(42), 17135-17140.
19. Vieider, F. M., Lefebvre, M., Bouchouicha, R., Chmura, T., Hakimov, R., Krawczyk, M., & Martinsson, P. (2015). Common components of risk and uncertainty attitudes across contexts and domains: Evidence from 30 countries. Journal of the European Economic Association, 13(3), 421-452.
20. Zethraeus, N., Kocoska-Maras, L., Ellingsen, T., Von Schoultz, B. O., Hirschberg, A. L., & Johannesson, M. (2009). A randomized trial of the effect of estrogen and testosterone on economic behavior. Proceedings of the National Academy of Sciences, 106(16), 6535-6538.