MANAGEMENT SCIENCE
Vol. 56, No. 11, November 2010, pp. 1977–1996
ISSN 0025-1909 | EISSN 1526-5501 | 10 | 5611 | 1977

informs®

DOI 10.1287/mnsc.1100.1226
© 2010 INFORMS

Downloaded from informs.org by [132.174.255.116] on 14 November 2013, at 06:45 . For personal use only, all rights reserved.

Prediction Markets: Alternative Mechanisms for Complex Environments with Few Traders

Paul J. Healy
Department of Economics, The Ohio State University, Columbus, Ohio 43210, healy.52@osu.edu

Sera Linardi
Division of the Humanities and Social Sciences, California Institute of Technology, Pasadena, California 91125, slinardi@hss.caltech.edu

J. Richard Lowery
Finance Department, McCombs School of Business, The University of Texas at Austin, Austin, Texas 78712, richard.lowery@mccombs.utexas.edu

John O. Ledyard
Division of the Humanities and Social Sciences, California Institute of Technology, Pasadena, California 91125, jledyard@hss.catlech.edu

Double auction prediction markets have proven successful in large-scale applications such as elections and sporting events. Consequently, several large corporations have adopted these markets for smaller-scale internal applications where information may be complex and the number of traders is small. Using laboratory experiments, we test the performance of the double auction in complex environments with few traders and compare it to three alternative mechanisms. When information is complex we find that an iterated poll (or Delphi method) outperforms the double auction mechanism. We present five behavioral observations that may explain why the poll performs better in these settings.

Key words: information aggregation; prediction markets; mechanism design
History: Received July 23, 2009; accepted June 14, 2010, by Teck-Hua Ho, decision analysis. Published online in *Articles in Advance* October 11, 2010.

1. Introduction

In large-scale applications, double auction prediction markets have proven successful at predicting future outcomes. The Iowa Electronic Market and the TradeSports–InTrade exchanges have outperformed national polls in predicting winners of political elections (Berg et al. 2008, Wolfers and Zitzewitz 2004), as did an underground political betting market in the late nineteenth and early twentieth centuries (Rhode and Strumpf 2004). Even markets with "play" money incentives such as the Hollywood Stock Exchange and the NewsFutures World News Exchange perform as well as real-money exchanges in predictive accuracy (Servan-Schreiber et al. 2004, Rosenbloom and Notz 2006).[1]

These successes in large-scale applications have led many large corporations—including Google, Hewlett-Packard, and Intel—to adopt standard double auction prediction markets for smaller-scale internal applications such as predicting future sales volumes of a particular product (Plott and Chen 2002, Hopman 2007, Cowgill et al. 2009).[2] It is not obvious, however, that the successes observed in large-scale settings will extend to most applications within corporations. Corporate prediction markets will involve far fewer traders, and they are likely to be used to address far more complex problems than those addressed in the relatively simple environments where the double auction mechanism has performed well. Management may want to collect information on variables that are correlated along several dimensions, such as demand for related products or costs across production units. Although standard double auction markets should be capable of aggregating this information in theory, it may be difficult in practice when traders face cognitive constraints and uncertainty about the rationality of others. These problems are exacerbated by the use of a relatively small number of traders because

[1] Rosenbloom and Notz (2006) do find that TradeSports significantly outperforms NewsFutures for some bundles of commodities and with enough data, but most tests cannot reject the null hypothesis of equal accuracy.

[2] Cowgill et al. (2009) identify at least 21 sizeable corporations that have used prediction mechanisms.

RIGHTSLINK

Downloaded from informs.org by [132.174.255.116] on 14 November 2013, at 06:45 . For personal use only, all rights reserved.

individuals may have market power that prevents convergence to the perfectly competitive outcome and therefore hinders the potential for information aggregation. In short, the assumptions of rational expectations and perfectly competitive markets seem at odds with the corporate environments where these markets are now being applied.

Given these potential difficulties there may be alternative information aggregation mechanisms that would outperform the standard double auction prediction market in smaller-scale settings with complex or dispersed information. For example, a variant of the Delphi method—where informed parties make predictions, learn each others' predictions, and then revise their own predictions—could be used to aggregate individuals' beliefs or private information, or a pari-mutuel-style betting market could be run to estimate the odds of certain future events.

In this paper, we employ a behavioral mechanism design methodology, using laboratory experiments to test the performance of the double auction mechanism in environments with a small number of traders (we use groups of only *three* traders in each mechanism) and complex information structures. We extend our analysis by comparing market performance in an environment with a moderately complex information structure with only one true-false event to a second environment with a highly complex information structure featuring three correlated true-false events. We then compare the double auction market's performance in these environments to the performances of three alternative mechanisms for aggregating information. Specifically, we compare the standard double auction mechanism to an iterated polling mechanism, a pari-mutuel betting mechanism, and a synthetic "market scoring rule" developed by Hanson (2003). By exploring the performance of these mechanisms in the laboratory we can gain an understanding about the domains on which each succeeds or fails and we can also acquire some insight into the reasons *why* some mechanisms outperform others by understanding how agents' behavior is affected by the details of the mechanism. Ultimately, insights such as these serve as inputs into the "behavioral" mechanism design process, providing guidance to practitioners hoping to design information aggregation mechanisms for use in these complex and small-scale settings.

Our choice of three participants per market serves to represent situations where thin markets, strategic interactions, and informationally large traders are significant concerns. Even relatively small, real-world applications would likely operate with more than three traders, but such markets face a wide set of other complications that do not arise in the lab but could also contribute to these problems. Additionally, because it is well established that double auction markets perform well when there are many informationally small traders, the use of an extremely small market allows us to evaluate whether there is some point below which the standard double auction prediction market breaks down and is surpassed by an alternative mechanism.

We find that the double auction market mechanism performs relatively well in an environment with a simple information structure involving one true-false event. In contrast, when the information structure becomes complex—with three correlated events and eight securities—the iterative poll performs the best and the standard double auction the worst. Thus, we find strong support for the claim that the complexity of the environment interacts with the details of the mechanism. For example, traders in the double auction with eight securities tend to focus attention on a small subset of the eight markets, causing severe mispricing in the remaining markets. The iterated poll avoids this issue by requiring players to announce beliefs about all eight states of the world simultaneously. In this way the design of the mechanism can be used to overcome natural behavioral biases that hinder information aggregation.

Our results suggest the following guidance for practitioners: In simple settings with a large number of traders relative to the number of items being predicted, we suggest using the standard double auction mechanism. When the number of items being predicted is large, when the predicted events may be correlated, or when the number of traders is small, we suggest the incentivized iterated poll instead. For example, a highly specialized firm seeking to project sales of its primary product should use a standard double auction, even in the face of concerns about limited participation and strategic trading. A more diversified firm seeking to evaluate expected sales for potentially complementary (or substitutable) products should consider an iterative polling mechanism instead, particularly when the number of informed traders is small. One downside of the iterated poll is that it requires subsidy payments from the institution running the mechanism; the size of these subsidies is limited, however, because we suggest using this mechanism only when the number of traders is relatively small. For larger environments the unsubsidized double auction mechanism is preferable. The pari-mutuel mechanism is less desirable because it appears to suffer from no-trade outcomes where agents prefer to opt out of the mechanism entirely, as is predicted by the no-trade theorem of Milgrom and Stokey (1982). We do not suggest the market scoring rule (MSR) because it tends to suffer from informational "mirages" where the mechanism leans toward completely *incorrect* predictions.

RIGHTSLINK

Downloaded from informs.org by [132.174.255.116] on 14 November 2013, at 06:45 . For personal use only, all rights reserved.

Given that our experiment represents a "stress test" using only three traders, we demonstrate the possibility that the performance of the double auction mechanism can be dominated by alternative mechanisms. The exact conditions under which the double auction outperforms the poll (or vice versa) are not known, and we hesitate to recommend the use of laboratory experiments to test such a question because fine details of the real-world environment that are absent in the lab would blur such conclusions. Our recommendations are a bit more coarse; managers should use the double auction mechanism to answer simple questions about broad, aggregate measures of performance or the likelihood of success of individual projects or products, whereas they will be better served by using the iterated poll when detailed information is needed about complicated and interrelated outcomes, like sales of correlated products or relative performance across divisions.

We follow our main results on mechanism performance with an analysis of five behavioral observations that we believe are related to the failure of the market mechanism and the success of the iterated poll in the complex setting. First, we see several apparent attempts at market manipulation in the double auction mechanism and in the pari-mutuel, but very few in the iterated poll and MSR. This is expected in the iterated poll; all players receive the same earnings and therefore have no clear incentive to manipulate their opponents' information. Second, total payments in the poll and MSR are subsidized by the mechanism designer, so all traders have an incentive to participate actively. Third, traders in the market appear to focus attention on only a subset of the securities—a heuristic that is impossible in the poll because it requires each trader to submit an entire probability distribution. Finally, an aberrant or confused trader can significantly affect final outcomes in the market, pari-mutuel, or MSR, but not in the poll because the poll takes the average of traders' reports as the predictive distribution.

These five observations allow us to extrapolate our results beyond the four mechanisms tested and to guide the design of future mechanisms. For example, a designer of other mechanisms for information aggregation should consider those with aligned incentives, subsidized total payments (if feasible), a focus on entire probability distributions, and minimal reliance on any one individual's report. Our results also inform economic theory: Theories of market equilibration should take into account the tendency for traders to manipulate others or to focus attention (or coordinate) on a subset of available markets. As such theories are developed and refined they could then be used to inform the design of additional mechanisms.

This paper extends past work on market efficiency and information aggregation. The number of traders in the market is often cited as a factor that affects the degree of efficiency and information aggregation, though the effect likely depends on the proportion of traders who hold valuable information. Clearly, the presence of additional informed traders increases the amount of information that is available to aggregate, but the effect of additional noise traders with no private information is unclear. DeLong et al. (1990) argue that noise traders' uninformed trades can reduce the informational content of market prices and damage market efficiency, whereas Kyle (1985) shows how noise traders can provide profit opportunities for informed traders, inducing them to make larger trades and invest more resources—physical or cognitive—in the acquisition and integration of information. Empirical evidence on the issue is mixed; volume is positively correlated with accuracy in the Iowa Electronic Markets (Berg et al. 2008) but also leads to more pricing anomalies and slower convergence to terminal cash flows in TradeSports markets (Tetlock 2008). Experimental results are similarly mixed: Bloomfield et al. (2009) observe lower informational efficiency in the presence of uninformed traders whereas Joel Grus and John Ledyard (see Ledyard 2005) observe greater aggregation when an automated noise trader is present.

A second set of factors affecting information aggregation concerns the complexity of the information and dividend structures in the market. These issues are amenable to laboratory studies given the difficulty in observing and controlling private information in field settings. Early experimental studies by Plott and Sunder (1988) find convergence and efficiency if simple Arrow-Debreu securities are used that pay a fixed dividend if and only if their associated state occurs, the structure of private information is relatively simple (agents are told which state is *not* true), and there is no aggregate uncertainty (combining all private signals reveals the true state perfectly). This result is replicated for a 10-state environment with less informative private signals (draws from an urn) and aggregate uncertainty by Plott (2000); however, this replication uses approximately 90 subjects whereas the earlier laboratory experiments typically include around 12 or 16 subjects. Markets with more complicated "tiered" securities (where dividend payments are state dependent and vary in magnitude across trader types) generate mixed results; having some traders know the state of the world perfectly, common knowledge of the dividend structures for all types, market experience, and a small number of tiered securities all facilitate convergence and efficiency (Plott and Sunder 1982, 1988; Forsythe and Lundholm 1990; O'Brien and Srivastava 1991).

RIGHTSLINK

Downloaded from informs.org by [132.174.255.116] on 14 November 2013, at 06:45 . For personal use only, all rights reserved.

From 2001 to 2003, John Ledyard, Robin Hanson, David Porter, and others worked to implement a prediction mechanism to forecast political and economic instability in the Middle East (see Hanson 2007 for details). The state space for this application becomes prohibitively large for any reasonable question of interest; if one wants to predict which of eight countries will experience GDP growth next quarter then $2^8 = 256$ separate securities are needed to capture the possibility that the likelihood of growth in each country depends on growth in the others. Unless the number of traders is large, the simple act of equilibrating all 256 markets (even with complete information) seems overwhelming.[3] In this complex environment (Ledyard et al. 2009), test the performance of a double auction that uses only 8 states—effectively ignoring the cross-country correlations—against five other mechanisms that used all 256 states: a combinatorial call market that allowed for trading of events like "X and Y" or "X given Y"; an individual proper scoring rule; a linear opinion pool; a logarithmic opinion pool; and the MSR developed by Hanson (2003), which is described below. Using groups of six subjects, the MSR and the opinion pools gave predictions closest to the full-information posterior. The eight-state double auction performed the worst, at least partially because they were necessarily handicapped by their inability to capture cross-country correlations. In a simpler environment with $2^3 = 8$ states and only three traders, the MSR is uniquely the best mechanism.

The current paper follows the work of Ledyard et al. (2009): We compare the double auction mechanism to three other mechanisms—an iterated poll, the pari-mutuel mechanism, and the MSR—in a relatively simple environment with only two states and a complex environment with $2^3 = 8$ states, each with only *three* traders per group. The latter environment is sufficiently large relative to the number of traders that we expect equilibration to be hindered by market liquidity shortages and subjects' cognitive limitations, but not so large that a simplification of the state space is necessary for the mechanism to operate.

Past studies have examined each of the mechanisms we test in different environments. McKelvey and Page (1990) study an iterated poll where each individual is paid on the accuracy of their own reports instead of the accuracy of the average report. This iterated poll fully aggregates all private information in theory but falls somewhat short of that target in the laboratory. Chen et al. (2001) also show how a poll outperforms a repeated call market with Arrow-Debreu securities as well as the information of the best-informed individual.[4] The pari-mutuel mechanism—used widely in horse race wagering—has similar theoretical properties to the double auction market: information should fully aggregate if trade occurs, but fully rational risk-averse traders should never have an incentive to trade. Plott et al. (2003) find that "prices" converge to the rational expectations prediction in a simple environment, but a simple model of trading based on private information alone predicts behavior better in more complex settings. In the field, Thaler and Ziembda (1988) show that pari-mutuels do a reasonably good job of predicting horse racing outcomes, though betters tend to over bet the unlikely ("long shot") horses.[5] Theoretically, the MSR fully aggregates information if traders are risk averse and *not forward looking*, but does provide some incentives for traders to misrepresent their information early to take advantage of others' incorrect beliefs later (see Chen et al. 2007 and Sami and Nikolova 2007 for two analyses of this mechanism). To our knowledge, only Ledyard et al. (2009)—who find that the MSR performs the best among their mechanisms—and this paper have tested the MSR in the laboratory.

We formally introduce the environments and mechanisms used in our study in §2. Section 3 details the experimental design. Results appear in §4, followed by analyses of our five observations in §5. We conclude with a discussion in §6.

2. Environments and Mechanisms

We consider an information aggregation problem where the state of the world consists of two dimensions. The first dimension represents some unobservable factor whose value impacts the realization in the second dimension. For example, the underlying monetary policy of a central bank (the first dimension) will affect whether or not the bank chooses to raise interest rates each quarter (the second dimension). Monetary policy is not directly observable, but interest rate movements are. In this setting traders in a double auction can use the bank's past interest rate changes to infer its monetary policy and, in turn, predict upcoming interest rate movements. If a collection of traders have different information about past

[3] Another concern is market manipulation by traders with an interest in the prediction generated by the market. Hanson et al. (2006) show in an experiment, however, that the accuracy of outside observers who use market prices to make predictions is not affected by the presence of these biased traders; Hanson and Oprea (2009) confirm theoretically that manipulators may play the same role as noise traders in Kyle (1985) and will therefore increase market efficiency.

[4] Chen et al. (2001) also adjust the aggregation of individual reports into a single posterior to account for subjects' risk aversion, though their adjustment does not significantly improve accuracy.

[5] Camerer (1998) attempts to manipulate actual horse races by placing and canceling large wagers, but the bettors return the odds to the "correct" values relatively quickly. Thus, the effects of manipulations are short-lived.

RIGHTSLINK

Downloaded from informs.org by [132.174.255.116] on 14 November 2013, at 06:45 . For personal use only, all rights reserved.

interest rate movements (and the underlying conditions of the economy at the time of those movements) then a double auction or other information aggregation mechanism can be used to generate more reliable predictions about the probability of future rate increases.

In the laboratory environment, we represent this inference problem by choosing one of two biased coins (the underlying first dimension) and then flipping the chosen coin one time (the second dimension that agents try to predict). The goal of an information aggregation mechanism is to predict the probability that the flip will land "heads." Subjects privately observe sample flips of the chosen coin, try to infer which biased coin was chosen, and then predict the probability that the one "true" flip will be heads. The goal of the mechanism designer is to combine these individual predictions into one aggregated prediction that incorporates all subjects' private information.[6]

Formally, the unknown true state of the world in our experimental environment is given by $(\theta, \omega) \in \Theta \times \Omega$ where θ (the coin) is drawn according to the distribution $f(\theta)$ and ω (the outcome of the coin flip) is drawn according to the conditional distribution $f(\omega \mid \theta)$. Each agent $i \in I$ privately observes K_i signals (sample coin flips) of ω, which we denote by $\hat{\omega}^i = (\hat{\omega}^i_1, \ldots, \hat{\omega}^i_{K_i}) \in \Omega^{K_i}$. Each $\hat{\omega}^i_k$ is drawn according to $f(\omega \mid \theta)$, so signals provide independent, unbiased information about θ that can then be used to predict the *true* value of ω.

Given the signal $\hat{\omega}^i$ and the priors $f(\theta)$ and $f(\omega \mid \theta)$, agent i forms a posterior belief $q(\theta \mid \hat{\omega}^i)$ over Θ using Bayes's rule. For simplicity, we denote this posterior on Θ by $q^i(\theta)$. From this, i forms a posterior over Ω given by $p^i(\omega) = \sum_{\theta' \in \Theta} f(\omega \mid \theta') q^i(\theta')$.

The goal of the mechanism designer is to aggregate the beliefs of the individual agents. The most accurate posterior the designer could hold in this setting would be that which she would form if she had *full information*, meaning she observes every agent's private signal. Letting $\hat{\omega} = (\hat{\omega}^1, \ldots, \hat{\omega}^I)$, we define $q^F(\theta) := q(\theta \mid \hat{\omega})$, which leads to the *full-information posterior* on Ω given by

$$p^F(\omega) = \sum_{\theta' \in \Theta} f(\omega \mid \theta') q^F(\theta').$$

To evaluate the performance of a given mechanism, we compare the belief distribution over Ω implied by behavior in the mechanism to the full-information posterior p^F. Abstracting away from the details, we think of mechanisms as producing a sequence of distributions over Ω denoted by $\{h_t\}_{t=0}^T$. Each distribution h_t represents the posterior at time $t \in \{0, \ldots, T\}$ implied by the messages sent by the players up through time t. Thus, h_0 corresponds to the prior, and we refer to h_T as the *output distribution* of the mechanism. At any point t, we call h_t the *running posterior* at time t. After observing the mechanism, the mechanism designer takes h_T as his posterior over Ω. *Full-information aggregation* occurs whenever the mechanism produces an output distribution equal to the full-information posterior, or $h_T \equiv p^F$. When Ω is finite we can measure the "error" of the output distribution, relative to the full-information posterior, by the normalized Euclidean norm[7]

$$\|h_T, p^F\|_\rho := |\Omega|^{1/2} \left(\sum_{\omega \in \Omega} |h_T(\omega) - p^F(\omega)|^2 \right)^{1/2}. \quad (1)$$

Our primary measure of the success of a mechanism is the average (or expected) size of this distance.

2.1. Environments

In our experiments we compare two environments that vary in the size of the state space and complexity of the information structure. The simpler environment is described above; one of two biased coins are chosen, and, upon flipping, the chosen coin either comes up heads or tails. Because there are two flip outcomes, we refer to this as the "two-state" environment. In the more complex environment, three biased and correlated coins are randomly ordered and then all three are flipped in the chosen order. There are eight possible outcomes of the flip of three coins, so we refer to this as the "eight-state" environment.[8] The two environments are described formally below. Recall that in both environments we use only three traders.

2.1.1. Two-State Environment. In the two-state design, $\Theta = \{X, Y\}$ and $\Omega = \{H, T\}$ with $f(\theta)$ and $f(\omega \mid \theta)$ given in Table 1. The interpretation is that one of two biased coins (X or Y) is to be randomly selected and flipped one time. The X coin is selected with probability 1/3 and comes up heads with probability 0.20. The Y coin is selected with probability 2/3 and comes up heads with probability 0.40. Agents observe neither the chosen coin (θ) nor the outcome of the flip (ω); instead, each agent observes sample flips of the chosen coin ($\hat{\omega}^i \in \Omega^{K_i}$), uses this information to form beliefs over which coin was selected (X or Y), and then forms a probability estimate that the one "true" coin flip is heads ($p^i(H)$).

[6] Our "sterile" version of the field setting allows us to test the ability of mechanisms to aggregate information in an (essentially) context-free environment. Our results therefore provide a baseline prediction about the relative performance of various mechanisms for use in any related field application.

[7] The normalization by $|\Omega|^{1/2}$ sets the norm of the centroid vector $(1/|\Omega|, \ldots, 1/|\Omega|)$ equal to one regardless of the size of Ω. This allows for casual comparison of distances between spaces of different dimension, though such comparisons should be made very cautiously.

[8] Technically, these names are misnomers because the true state spaces $(\Theta \times \Omega)$ are of size $2 \times 2 = 4$ and $6 \times 8 = 48$, respectively.

RIGHTSLINK

Downloaded from informs.org by [132.174.255.116] on 14 November 2013, at 06:45 . For personal use only, all rights reserved.

Table 1 Distribution f for the Two-State Experiments

θ	$f(\theta)$	$f(H \mid \theta)$	$f(T \mid \theta)$
X	1/3	0.2	0.8
Y	2/3	0.4	0.6

2.1.2. Eight-State Environment. In the eight-state design, there are three coins, X, Y, and Z, placed in a random order such as YZX or ZYX. The set Θ contains the six possible orderings, each of which is equally likely a priori. Once an ordering is chosen, the three coins are then flipped in that order. The result is a triple of heads and tails, such as HHT or THT, where the first character corresponds to the flip of the first coin in the order, the second character corresponds the second coin, and so on. The set Ω contains all eight possible flip outcomes. Agents do not know the true outcome of the flip of the three coins (ω) nor the actual ordering of the coins (θ); instead, they observe sample flips of the chosen coin ordering ($\hat{\omega}^i \in \Omega^{K_i}$), use this information to form beliefs over which of the six orderings was selected, and then form beliefs over the eight possible outcomes of the "true" coin flips ($p^i(HHT)$, $p^i(THT)$, etc.).

The X coin lands heads with probability 0.20 and the Z coin lands heads with probability 0.40. The Y coin is different; its flip matches the flip of the X coin with probability 2/3 and differs from X with probability 1/3. The values of $f(\theta)$ and $f(\omega \mid \theta)$ for this environment are given in Table 2.

Note that, unconditionally, the Y coin lands heads with probability 0.40, making it indistinguishable from the Z coin if one ignores the correlation between coins. In other words, an agent trying to infer the ordering of the three coins based on a sample of flips must first identify the X coin by its lower frequency of heads and then distinguish between the Y and Z coins by identifying which is correlated with X. When each agent has a small number of sample flips, this inference problem is difficult and the value of each agent's private information is small. This is the sense in which the eight-state environment is considered more complex.

One real-world setting with a similar correlation structure is the conference championship structure used in many professional and collegiate sports. Here, coin X represents the event that Team A beats Team B in the Western conference championship, coin Z represents the event that Team C beats Team D in the Eastern conference championship, and coin Y represents the event that the Western conference champion beats the Eastern conference champion in the final match-up. Clearly coin Y depends on which teams actually advance to the final game; thus, Y will be correlated with the other two coins. If probabilities were elicited for only the three games, this correlation would not be captured; it takes a full set of $2^3 = 8$ probabilities to capture this correlation.

Table 2 Distribution f for the Eight-State Experiments

θ	$f(\theta)$	TTT	TTH	THT	THH	HTT	HTH	HHT	HHH
XYZ	1/6	0.320	0.213	0.160	0.107	0.040	0.027	0.080	0.053
XZY	1/6	0.320	0.160	0.213	0.107	0.040	0.080	0.027	0.053
YXZ	1/6	0.320	0.213	0.040	0.027	0.160	0.107	0.080	0.053
YZX	1/6	0.320	0.040	0.213	0.027	0.160	0.080	0.107	0.053
ZXY	1/6	0.320	0.160	0.040	0.080	0.213	0.107	0.027	0.053
ZYX	1/6	0.320	0.040	0.160	0.080	0.213	0.027	0.107	0.053

2.2. Mechanisms

In any field application, a mechanism's performance—and, therefore, agents' payoffs—depends on the realized value of ω. Consequently, even mechanisms that fully aggregate information can perform poorly when an unlikely true state happens to occur. In the controlled laboratory setting, one way to reduce this variation is to reward subjects based on the *expected* performance of the mechanism given the true distribution $f(\omega \mid \theta)$.[9] In our experiments we generate an estimate of $f(\omega \mid \theta)$ using 500 draws of ω. Letting $\phi(\omega)$ denote the fraction of the 500 draws that equals ω, the empirical distribution ϕ serves as a close approximation to the true distribution $f(\omega \mid \theta)$.[10] Subjects are then paid based on the expected performance of the mechanism given the distribution $\phi(\omega)$. This is explained in more detail with each mechanism.

In what follows we index the elements of Ω by $s \in \{1, \dots, S\}$. In the two-state environment $S = 2$, and in the eight-state environment $S = 8$.

2.2.1. Double Auction. The standard prediction market mechanism used widely in field applications is a double auction with a complete set of Arrow-Debreu securities, henceforth referred to as the "double auction" mechanism. Here, S state-contingent securities (one for each $\omega_s \in \Omega$) are traded in separate markets. Subjects buy and sell each security in a standard computerized double auction format with an open book where all bids and asks are public information. Traders are initially endowed with cash but no assets; those who want to sell an asset do so by selling short and holding negative quantities. At the end of the trading period each asset s is worth $\phi(\omega_s)$ experimental dollars. Traders who own a positive quantity of asset s receive $\phi(\omega_s)$ experimental dollars per unit,

[9] This cannot be done in most field settings because θ is not observed.

[10] We chose to approximate $f(\omega \mid \theta)$ using $\phi(\omega)$ because the latter is constructed though an actual (computerized) process; we expect that this makes it more understandable to subjects without a statistics background.

Downloaded from informs.org by [132.174.255.116] on 14 November 2013, at 06:45 . For personal use only, all rights reserved.

and traders who hold a negative quantity of asset s pay $\phi(\omega_s)$ experimental dollars per unit.[11]

In a rational expectations equilibrium, the asset prices reveal all private information. Under certain assumptions about preferences, these equilibrium prices will equal the full-information posterior probabilities.[12] Thus, we set the mechanism output distribution equal to the vector of security prices. In our experiment the prices of the securities are not forced to sum to one, but in our data analysis we set all untraded security prices equal to the uniform distribution price of $1/|\Omega|$ and then proportionally adjust the prices of all traded securities so that the sum of all prices equals one. This generates a well-defined probability distribution as the mechanism's output.[13]

Because this mechanism is zero-sum, however, the no-trade theorem of Milgrom and Stokey (1982) implies that we should not expect any trade in equilibrium with risk-averse agents. Whether or not trade actually occurs and prices equilibrate to the full-information posterior, however, depends on the beliefs, preferences, and rationality of the traders.

2.2.2. Pari-Mutuel Betting. In pari-mutuel betting, traders buy "tickets" or "bets" on each of the S possible states. Tickets cost one experimental dollar each and a trader can buy as many tickets of each type as he can afford using his cash endowment. During the period, the total number of tickets of each type that have been purchased is displayed publicly. At the end of the period these totals are used to calculate the payoff odds for each security. If T_s is the total quantity of state-s tickets purchased then the payoff odds for state s are given by $O_s = (T_s/\sum_\omega T_\omega)^{-1}$. Each state-$s$ ticket is then redeemed for $O_s \cdot \phi(\omega_s)$ experimental dollars. In other words, each ticket is worth the payoff odds times the (approximated) true probability that state ω_s occurs.

The total payoff across all tickets and individuals equals the sum of all purchases, making this a zero-sum game. As in the double auction a no-trade theorem applies, so risk-averse agents should not purchase tickets in an equilibrium with common knowledge of rationality. In the presence of noise trading, however, rational traders may have an incentive to participate. It is certainly the case that, once information has fully aggregated, rational, risk-averse agents will purchase tickets to move the inverse of the payoff odds to the (common) posterior probabilities. In other words, the fraction of total tickets outstanding that are state-s tickets should equal the state-s posterior probability. For this reason we set the mechanism output probability of each state ω equal to the fraction of total tickets outstanding that are state-ω tickets. Whether or not information will actually aggregate, however, is a question for the laboratory.

2.2.3. Iterative Polls. Iterative polls—an incentivized version of the "Delphi method"—are perhaps the simplest and most direct information aggregation mechanism. Subjects are asked to report simultaneously a probability distribution over Ω. The reports are averaged across subjects (by taking the arithmetic mean of the reports for each state) to generate an "aggregated" report. This aggregated report is shown to all subjects, who are then asked to submit simultaneously a second distribution over Ω. Subjects' second reports will incorporate their own private information plus any information inferred from the average of the first reports. The average of these second reports is displayed, and the process is repeated for a total of five reports. The fifth average report is then taken as the output distribution of the mechanism.

All subjects are all paid identically based on the accuracy of the final report using a logarithmic scoring rule. Specifically, if $h_T(\omega_s)$ is the final average probability report then for each state ω_s each subject i is given $\ln(h_T(\omega_s)) - \ln(1/S)$ tickets. Thus, agents gain state-s tickets if $h_T(\omega_s) > 1/S$ and lose state-s tickets if $h_T(\omega_s) < 1/S$. Once the empirical frequency ϕ is revealed each state-s ticket pays out $\phi(\omega_s)$ dollars. Because all agents receive the same payment, the game is not zero-sum and therefore must be subsidized by the mechanism designer.

The logarithmic scoring rule is incentive compatible (Selten 1998), so any risk-neutral individual acting in isolation would prefer to announce truthfully her beliefs over Ω. In the multiple-player game, there exist sequential equilibria in which full-information aggregation occurs; thus, we take the final average announcement to be the mechanism's output distribution. One might conjecture that *any* sequential equilibrium should feature full-information aggregation because all players have identical incentives, but in fact there exist "babbling" equilibria in which full-information aggregation does not occur.[14] Under risk neutrality the full-information aggregation equilibria

[11] In field applications the asset corresponding to the true state is worth one dollar and all other assets are worthless.

[12] See Manski (2006), Wolfers and Zitzewitz (2006), and Gjerstad (2005).

[13] We could, alternatively, set the prices of nontraded securities equal to zero when prices sum to more than one and then proportionately adjust the prices of the traded securities, while distributing the residual probability over the nontraded securities when the prices sum to less than one. This approach generates larger errors for the double auction, and under this alternative the double auction performs worse than all other mechanisms at a very high significance level.

[14] In a "good" equilibrium each player announces truthfully in the first round, all players use the first average report to infer others' information, then all players announce the full-information posterior in rounds two through five, ignoring any deviations by

RIGHTSLINK

Downloaded from informs.org by [132.174.255.116] on 14 November 2013, at 06:45 . For personal use only, all rights reserved.

are Pareto dominant, so the success of the mechanism depends on agents' ability to coordinate on this payoff-dominant outcome.

2.2.4. Market Scoring Rule. In the MSR, a probability distribution $h_0 = (h_0(\omega_1), \ldots, h_0(\omega_S))$ is publicly displayed at the beginning of each period; in our experiments, $h_0(\omega_s) = 1/S$ for each s. At any given time t during the period, any trader may "move" the current distribution h_t to a new distribution, h_{t+1}. This is done simply by announcing the new distribution h_{t+1}. When a trader makes such a move he receives (or loses)

$$\ln(h_{t+1}(\omega_s)) - \ln(h_t(\omega_s)) \qquad (2)$$

state-s tickets for each s. Traders are given an initial endowment of tickets and cannot move h_t to some h_{t+1} if such a move would require surrendering more tickets of some state than the trader currently holds. This prevents traders from moving probabilities arbitrarily close to zero because the logarithm becomes infinitely negative for arbitrarily small probabilities.

During the period, traders may move the probability distribution as many times as they like, subject to the budget constraint. With each move, they gain and lose tickets appropriately. At the end of the period each state-s ticket is worth $\phi(\omega_s)$ experimental dollars. Because summing Equation (2) over all t yields

$$\ln(h_T(\omega_s)) - \ln(h_0(\omega_s)),$$

the total change in ticket holdings depends only on the starting distribution h_0 and the ending distribution h_T (intuitively, each trader is "buying out" the position of the previous trader). The final cash value of this difference must be subsidized (or collected) by the mechanism designer.

As in the iterated poll, this mechanism uses the logarithmic scoring rule which is incentive compatible for any risk neutral individual, meaning players will truthfully reveal their beliefs if they do not expect to make any future moves. Thus, if it is common knowledge that each player's final move is in fact their last then each will fully reveal their beliefs in the final move and information will fully aggregate in the final move of the period.[15] For this reason, we take the final move of the period to be the output distribution of the mechanism.

If a player does expect to move again in the future then there may be an incentive to misrepresent one's information so that other players erroneously move the distribution away from the full-information posterior, and the misrepresenting player can then earn profits by moving it back. In our experiment players can make moves at any time during the five-minute window, so it is not clear whether manipulations will persist through the final move or whether information will fully aggregate at the end of the period. We test for evidence of manipulations in §4.

3. Experimental Design

All experiments were run at the California Institute of Technology using undergraduate students recruited via e-mail. Each period lasted five minutes, and subjects earned an average of approximately $30 per session.

In each period subjects are publicly informed about the distribution f given in Tables 1 and 2, so we take this as the common prior.[16] A coin (or coin ordering) θ is chosen by the computer but not revealed to the subjects. Instead, each subject is privately shown a unique sample of coin flips of the chosen coin. The mechanism is then run and the output distribution is observed. After the period ends traders are told the chosen coin and the distribution $\phi(\omega)$ generated from 500 sample flips of the chosen coin.[17] Subjects' total earnings are then augmented by their payment for the period and the next period begins.

Following the standard practice in experimental economics, the framing of this experiment is entirely neutral. States are described to subjects as "coins" and "coin flips." Real business contexts may alter performance somewhat, but the neutral frame can be taken as a "baseline" environment against which all context-laden settings can be compared. Based on past evidence, we expect the results from the neutral experiment to an unbiased predictor of real-world performance (see, e.g., Fréchette 2009).

A 4 × 2 experimental design compares the four mechanisms described in §2.2 in both the two-state and eight-state environments. Agents participate in groups of three and are matched with the same group for the entire experiment. Each subject group participates in one mechanism for eight periods followed by a different mechanism for eight periods. We use a crossover design in which the ordering of mechanisms for one group is then reversed for another group. Each

others. In a "babbling" equilibrium all players submit random, meaningless announcements in rounds one through four, ignore others' announcements, and attempt to maximize their payoff in the final round; because no information was conveyed in the first four rounds, the final average report generically will not achieve full-information aggregation.

[15] This argument is based on the analysis of Chen et al. (2007); see also Sami and Nikolova (2007).

[16] Technically, the prior is common *information* but not necessarily common knowledge.

[17] All individual signals are independent and independent of the 500 flips used to determine $\phi(\omega)$.

RIGHTSLINK

Downloaded from informs.org by [132.174.255.116] on 14 November 2013, at 06:45 . For personal use only, all rights reserved.

Table 3 Experimental Design

Session number	No. of states	No. of agents	Mechanism 1 (periods 1–8)	Mechanism 2 (periods 9–16)
1	2	3	Pari-mutuel	Market scoring rule
2	2	3	Pari-mutuel	Market scoring rule
3	2	3	Market scoring rule	Pari-mutuel
4	2	3	Market scoring rule	Pari-mutuel
5	2	3	Double auction	Iterative poll
6	2	3	Double auction	Iterative poll
7	2	3	Iterative poll	Double auction
8	2	3	Iterative poll	Double auction
9	8	3	Pari-mutuel	Market scoring rule
10	8	3	Pari-mutuel	Market scoring rule
11	8	3	Market scoring rule	Pari-mutuel
12	8	3	Market scoring rule	Pari-mutuel
13	8	3	Double auction	Iterative poll
14	8	3	Double auction	Iterative poll
15	8	3	Iterative poll	Double auction
16	8	3	Iterative poll	Double auction

ordering is run twice for a total of 16 experimental sessions.[18] Table 3 lists the details of each session.

The MSR, pari-mutuel, and poll were all run manually. Subjects sat at desks and a spreadsheet program was projected onto a screen at the front of the room. In the MSR and pari-mutuel, bids were submitted in a continuous-time, open-outcry manner. In each round of the poll, subjects privately and simultaneously submitted their announcements on paper. In all three mechanisms, the submitted bids or announcements were immediately entered into the spreadsheet and the current market prices were automatically updated on the screen. The double auction was run using the jMarkets software package. This software uses a visual interface, features an open book so all traders can see outstanding bids and offers, and allows continuous-time trading.

In each mechanism, after a period had ended, players were shown the distribution of "true" coin flips, their payoffs, and then given a slip of paper containing their private information for the following period. Subjects have access to standard calculators (but not payoff calculators specific to these mechanisms), pencil, and paper throughout the experiment.

4. Results

The results are organized as follows: First we describe the four ways in which we measure the performance (or failure) of each mechanism. We then show that behavior does not significantly differ across periods and does not depend on whether a mechanism is presented first or second within a given session, allowing us to aggregate results across periods and orderings and directly compare the four mechanisms using our four performance measures.

4.1. Measures of Performance

Our primary measure of a mechanism's performance is the average normalized Euclidean distance between the mechanism's output distribution h_T and the full-information posterior p^F (see Equation (1)); this provides a simple measure of how accurate the mechanism designer's posterior beliefs are relative to the ideal case of full-information aggregation.[19]

One might also be concerned with other properties of the mechanism's performance. For example, consider the no-trade theorem in the context of the double auction and pari-mutuel mechanisms. In a thin market devoid of noise traders, (weakly) risk-averse rational traders should (weakly) prefer not to participate in either mechanism. If no trade occurs then the mechanism provides no value to the designer because no new information is revealed. If the market were sufficiently thick then it becomes more likely that noise traders will exist—or at least that rational traders believe that noise traders exist—and so trade will occur and information will be revealed. In our experiments, however, groups contain only three agents so the logic of the no-trade theorem is particularly compelling in this setting.

Worse than the no-trade outcome is a situation where the mechanism output is misleading. For example, if a mechanism's output distribution in the two-state environment indicates that heads is *less* likely than previously expected when in fact the private information indicates that heads is *more* likely to occur then the designer's posterior is less accurate than the prior. This outcome has been called a *mirage* in the existing literature (Camerer and Weigelt 1991). In general, we label an output distribution as a mirage if it lies in the opposite direction from the prior as the full-information posterior. Formally, a mirage occurs when $(p^{FI} - p_0) \cdot (h_T - p_0) < 0$, where p^0 is the prior, h_T is the output distribution, and p^{FI} is the full-information posterior. Graphical representations of a mirage (for both two- and eight-state environments) are provided in Figure 1.

A third possible failure of a mechanism is a situation where the output distribution cannot be rationalized by Bayes's rule. We label such outcomes as *Bayes-inconsistent*. For example, the probability of heads in the two-state environment must lie between 0.2 (the probability of heads for the X coin) and 0.4 (the

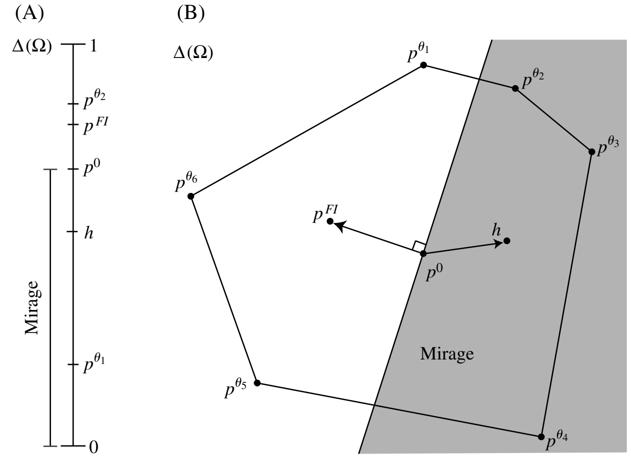
[18] We pair the pari-mutuel with the MSR and the double auction with the poll. This choice is arbitrary; what matters is that for each pairing we run both orderings of that pairing to test for ordering or learning effects.

[19] Other distance measures such as the Kullback and Leibler (1951) information criterion generate qualitatively similar results.

RIGHTSLINK

Downloaded from informs.org by [132.174.255.116] on 14 November 2013, at 06:45 . For personal use only, all rights reserved.

Figure 1 Mirages with (A) Two States and (B) More Than Two States



Notes. In panel (A), the mechanism output h lies between the prior p^0 and the probability associated with state θ_1, whereas the posterior implied by all private information (p^{FI}) lies between the prior and the probability associated with state θ_2. In panel (B), the full-information posterior p^{FI} implies that states θ_1, θ_5, and θ_6 are relatively more likely than under the prior, whereas the mechanism output h would lead to the conclusion that states θ_2, θ_3, and θ_4 are more likely.

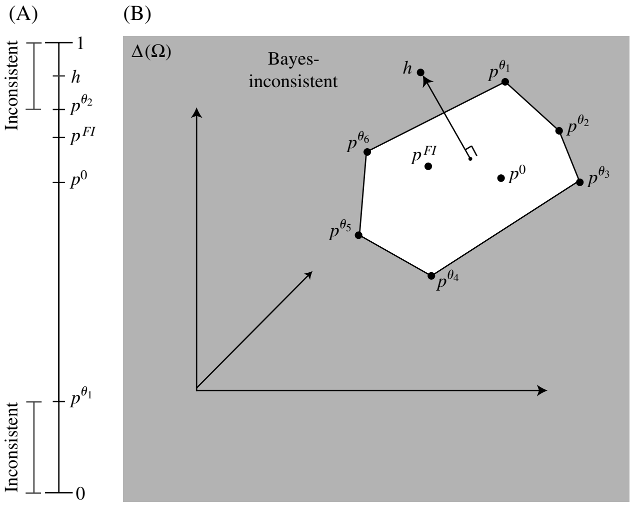
probability of heads for the Y coin).[20] If the mechanism output probability of heads is 0.43 then the logic of standard probability theory offers no advice as to what the best prediction should be; certainly one could construct ad hoc theories to rationalize this output and generate a prediction, but from our view this output represents a failure of the mechanism precisely because such ad hoc theories become necessary. Graphical representations of Bayes-inconsistent outcomes (for two and eight states) is provided in Figure 2.

For each mechanism in each environment, we compare the distance to the full-information posterior and count the number of periods in which no trading, mirages, or inconsistencies occur.[21]

4.2. Period and Order Effects

Although one might expect learning and experience to generate better performance in later periods, we do not find strong evidence for this hypothesis. Using a Wilcoxon rank sum test, we compare the distance between the mechanism output distribution and the full-information posterior for each period t against the distance for each period $s \neq t$. Aggregating across all four mechanisms, we cannot reject the hypothesis that the distances have equal distributions for any pair of periods in the two-state experiments or in the eight-state experiments. Thus, for example, the distribution of first-period distances is approximately the same as the distribution of last-period distances, indicating that no significant learning takes place. This is clear from panels (A) and (B) of Figure 3.[22] The same set of tests run on each mechanism (rather than aggregating across all four mechanisms) generates the same results.[23]

Figure 2 Bayes-Inconsistent Outcomes with (A) Two States and (B) More Than Two States



Notes. In panel (A), θ_2 represents the state where the probability of the outcome in question is highest, but the mechanism output implies a posterior probability higher than the probability if θ_2 was known for certain to be the state. In panel (B), the mechanism output implies a posterior probability that cannot be rationalized by any belief about the underlying state because the outcome lies outside the convex hull of probabilities implied by each state.

Because subjects participate in one mechanism for eight periods and then a second mechanism for a subsequent eight periods, some experience from the first mechanism may spill over into the second mechanism, creating a mechanism ordering effect in our data. Comparing the distance between the mechanism output and the full-information posterior for mechanisms run in the first eight periods versus those run in the final eight periods reveals no discernible effect; aggregating across all four mechanisms, Wilcoxon tests find no significant difference for both the two-state experiments ($p = 0.820$) and the eight-state experiments ($p = 0.850$). The same tests run on each mechanism individually also find no significant effect (all p-values are

[20] For a formal proof of this fact more generally, see Shmaya and Yariv (2008).

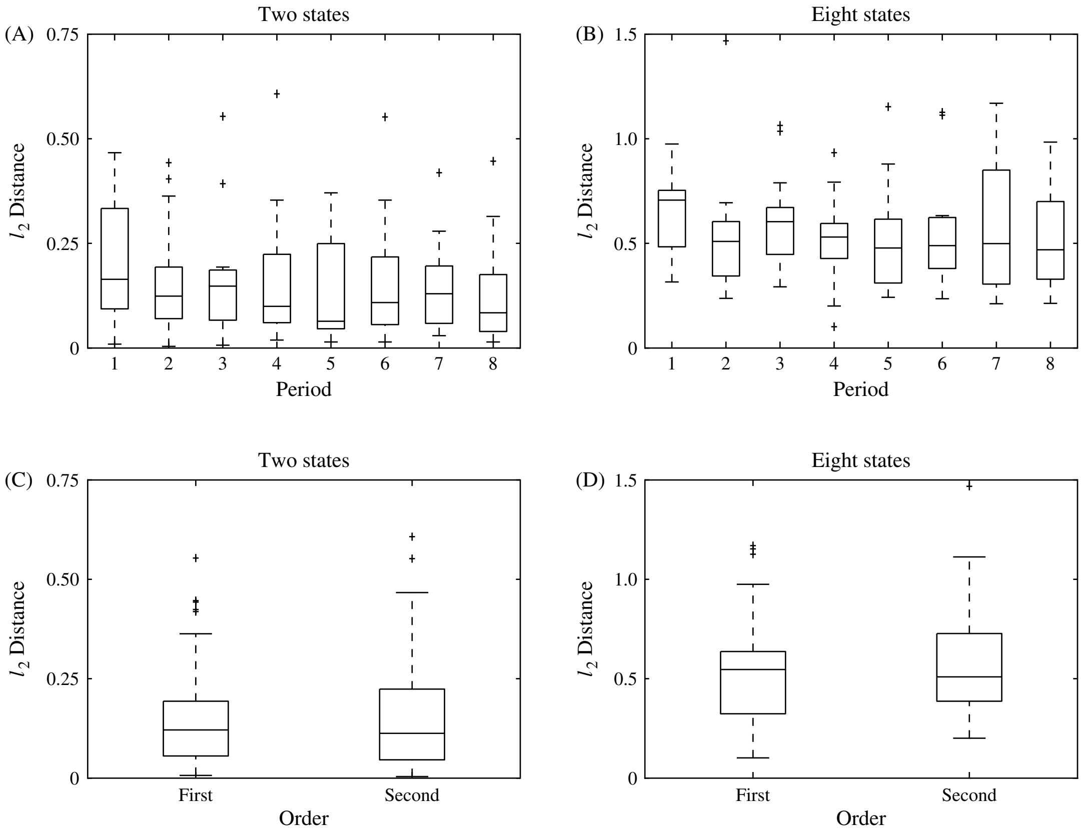
[21] We have also constructed various measures of the degree to which each failure occurs; these results are qualitatively similar to counting the number of failures

[22] The two- and eight-state figures are scaled differently to maximize readability; recall that comparisons of errors across these cases are not meaningful.

[23] Specifically, of the 112 period-versus-period tests, we find that four (or 3.6% of the tests) are significant at the 5% level in the two-state experiments and none are significant at the 5% level in the eight-state experiments.

Downloaded from informs.org by [132.174.255.116] on 14 November 2013, at 06:45 . For personal use only, all rights reserved.

Figure 3 Box-and-Whisker Plots of the Distance Between the Mechanism Output Distribution and the Full-Information Posterior for (A) Each Period in the Two-State Experiments, (B) Each Period in the Eight-State Experiments, (C) Each Mechanism Ordering in the Two-State Experiments, and (D) Each Mechanism Ordering in the Eight-State Experiments



greater than 0.168). The plots in panels (C) and (D) of Figure 3 demonstrate this result.

Because we find no significant period or ordering effects, we aggregate across all periods and both orderings in all subsequent analyses.

4.3. The Simple Environment: Two States

4.3.1. Mechanism Accuracy. To determine which mechanisms are the most accurate, we perform a comparison of the mechanism error (distance from the mechanism output to the full-information posterior) between each pair of mechanisms.[24] For every given pair, we aggregate across all periods and orderings from the two-state experiments and perform a Wilcoxon test on the resulting distributions of errors. From these comparisons, we can construct a "significance relation" that ranks the four mechanisms according to the degree of error they generate.

Formally, we define the significance relation by $A \succeq B$ if mechanism A has a higher average error than B and $A \succ B$ if that difference is statistically significant at the 10% level. Because $\succ$ is not negatively transitive (it is possible to have $A \not\succ B$ and $B \not\succ C$ but $A \succ C$), describing the relation between mechanisms may require multiple statements. For example, from the pair of statements $A \succeq B \succeq C \succeq D$ and $A \succ C \succeq D$, we conclude that A has significantly higher average error than C and D, but that A's average error is not significantly greater than B's and that no other comparisons are statistically significant.

The result of the pairwise comparison procedure is reported in Table 4, and the distributions of errors for each mechanism are shown in panel (A) of Figure 4. The average error for each mechanism is reported in the second row and second column of the table; on average the MSR generates the largest errors and the double auction generates the smallest errors. The p-values of the pairwise Wilcoxon tests are reported in the third through sixth columns and the third through sixth rows. No differences are significant at the 5% level, but the market scoring rule generates significantly higher error than both the poll and the double auction at the 10% level. From this, we generate the

[24] Because we use a distance measure, we do not separate error caused by systematic bias and error caused by noise. In separate tests for the simple environment, we do not reject the null hypothesis that the average *signed* error is zero for each mechanism, indicating no systematic bias in the mechanisms' output distributions.

Downloaded from informs.org by [132.174.255.116] on 14 November 2013, at 06:45 . For personal use only, all rights reserved.

Table 4 *p*-Values of Mechanism-by-Mechanism Wilcoxon Tests on the Distance to the Full-Information Posterior for the Two-State Experiments

Two states	Avg. distance	Dbl. auction	Mkt. scoring rule	Pari-mutuel	Poll
Avg. distance	—	0.131	0.210	0.148	0.133
Dbl. auction	0.131	—	*0.092*	0.646	0.663
Mkt. scoring rule	0.210	—	—	0.225	*0.098*
Pari-mutuel	0.148	—	—	—	0.519
Poll	0.133	—	—	—	—

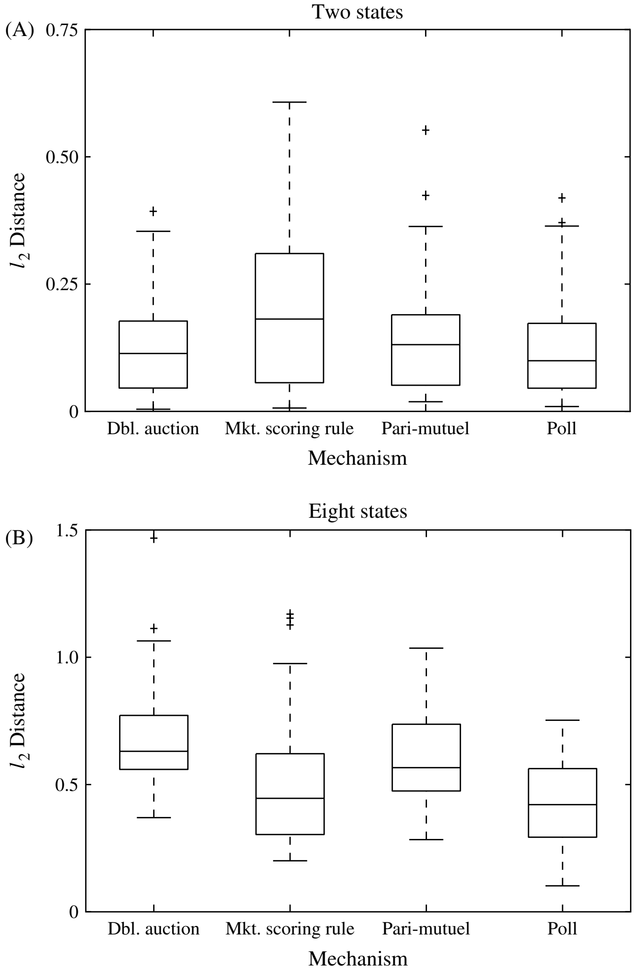
Notes. 10% Significance ordering: MSR $\succeq$ Pari $\succeq$ Poll $\succeq$ DblAuc and MSR $\succ$ Poll $\succeq$ DblAuc. Italicized entries are significant at the 10% level.

significance statements: "MSR $\succeq$ Pari $\succeq$ Poll $\succeq$ DblAuc and MSR $\succ$ Poll $\succeq$ DblAuc." Thus, the MSR is the only mechanism that generates significantly higher error than any other mechanism. In other words, these results are not particularly conclusive about which mechanism is the best (in terms of error), but the results are clear about which mechanism is the worst.

The Wilcoxon tests in Table 4 treat each period in each session as an independent observation, potentially biasing the results if cohort effects are present. Using Wilcoxon tests to compare the error measures from each pair of sessions in each mechanism, we find little evidence of cohort effects: 2 of 24 session-pairs have significant differences at the 10% level (one in the double auction and one in the poll). This is roughly the number of significant differences one should expect under the null hypothesis of no cohort effects, so we do not reject that hypothesis. Clearly, if one were to treat each session as a single observation, the marginally significant comparisons in Table 4 would become insignificant.

If observations within a cohort can be viewed as independent (which may be valid because no period effects are found), controlling for cohorts can strengthen the comparison between mechanisms. For example, an ANOVA analysis treating cohorts as a nested factor within each mechanism removes the between-cohort variability from the error data.[25] With this extra statistical power the marginally significant results in Table 4 ("MSR $\succ$ Poll" and "MSR $\succ$ DblAuc") become significant at the 5% level (*p*-values of 0.022 and 0.026, respectively). None of the other comparisons becomes significant at the 10% level. Thus, we strengthen our conclusion that the MSR generates the largest errors in the simple environment.

Figure 4 Box-and-Whisker Plots of the Distance Between the Mechanism Output Distribution and the Full-Information Posterior for Each Mechanism in (A) the Two-State Experiments, and (B) the Eight-State Experiments



4.3.2. Catastrophes: No Trade. In theory, we predict no trade (or indifference to trade) in the double auction and pari-mutuel mechanisms when agents are (weakly) risk averse. In practice (see the second row of Table 5), we observe trade in all 32 periods of the double auction, but no trade in 4 of the 32 periods (12.5%) of the pari-mutuel mechanism, all in Session 3. Despite the fact that it is subsidized—thus circumventing the no-trade issue in theory—we do observe one period of no trade in the MSR. Because all instances of no trade occur in a single session for both mechanism, we cannot disentangle mechanism effects from session/cohort effects and therefore cannot employ proper panel data techniques to compare the rate of no-trade between mechanisms. Using a simple two-sample binomial test (which incorrectly assumes independence of no-trade periods) as a rough

[25] Because there are no-trade periods, this becomes an unbalanced nested two-factor design. We test for pairwise mechanism effects by running dummy-variable regressions, comparing the error sum-of-squares of a full model with all mechanism and cohort dummies included to the error sum-of-squares of a restricted model where two mechanisms' effects are constrained to be equal. An *F*-test then determines whether the full model gains significant explanatory power over the restricted model, and therefore whether or not the true mechanism effects are equal. See Neter et al. (1996, pp. 1138–1141) for details. Diagnostics of residuals suggest that the required parametric assumptions are reasonably satisfied.

Downloaded from informs.org by [132.174.255.116] on 14 November 2013, at 06:45 . For personal use only, all rights reserved.

Table 5 Number of Periods in Each Session (Out of 8) and Number of Periods Total (Out of 32) in Which Each Type of Catastrophic Failure Occurs in the Two-State Experiments

	Dbl. auction		Mkt. scoring rule		Pari-mutuel		Poll	
	(S5, S6, S7, S8)	Tot.	(S3, S4, S1, S2)	Tot.	(S1, S2, S3, S4)	Tot.	(S7, S8, S5, S6)	Tot.
No trade	(0, 0, 0, 0)	0	(0, 1, 0, 0)	1	(0, 0, 4, 0)	4	(0, 0, 0, 0)	0
Mirage	(4, 4, 2, 3)	13	(3, 2, 3, 5)	13	(2, 4, 0, 3)	9	(2, 3, 2, 3)	10
Bayes-inconsistent	(2, 0, 2, 1)	5	(3, 1, 3, 0)	7	(2, 2, 2, 0)	6	(3, 3, 1, 4)	11
Bayes-inc. mirage	(0, 0, 0, 0)	0	(1, 0, 0, 0)	1	(0, 1, 0, 0)	1	(2, 1, 0, 0)	3
None	(2, 4, 4, 4)	14	(3, 4, 2, 3)	12	(4, 3, 2, 5)	14	(5, 3, 5, 1)	14

guide, we conclude that the pari-mutuel mechanism generates no-trade outcomes more frequently than the double auction and poll (both with one-tailed p-values of 0.034) but not the MSR (p-value: 0.118). We therefore suggest that the pari-mutuel is more vulnerable to no-trade than either the double auction or poll.

Intuitively, we conjecture that subjects are prone to trade, whether or not rational, in the more familiar double auction mechanism and are prone to confusion and, consequently, inactivity in the unfamiliar and mathematically complex market scoring rule mechanism. As for the pari-mutuel mechanism, debriefing discussions with subjects indicated that several believed that first movers would be disadvantaged in this zero-sum game because placing a wager may reveal valuable private information, allowing competitors to gain at the first mover's expense.[26]

4.3.3. Catastrophes: Mirages. The frequency of mirages for the two-state experiments is reported in the third row of Table 5. Although all four mechanisms generate a substantial frequency of mirages (ranging from 31% to 44%), the differences between mechanisms not statistically significant in either simple binomial tests or in a random effects probit model, which controls for cohort effects. Furthermore, several periods of the pari-mutuel and poll had output distributions equal to the prior; if these periods are also counted as mirages, the mechanisms perform very similarly by this measure (with 13, 14, 15, and 13 mirages, respectively).

4.3.4. Catastrophes: Inconsistencies. The fourth row of Table 5 displays the number of periods in which Bayes-inconsistent outcomes occur in the two-state experiments.[27] Clearly the poll is the most frequent; using a probit random effects model, we conclude that the poll generates Bayes-inconsistent outcomes significantly (at the 10% level) more frequently than the double auction (p-value of 0.084). Thus, our significance statement regarding Bayes-inconsistency is "Poll $\succeq$ MSR $\succeq$ Pari $\succeq$ DblAuc and Poll $\succ$ DlbAuc." Conditional on observing a Bayes-inconsistent outcome, the average distance between h and the convex hull ($[0.2, 0.4]$) is 0.024, 0.171, 0.106, and 0.052 for the double auction, MSR, pari-mutuel, and poll, respectively. Thus, the "magnitude" of the Bayes-inconsistency in the poll is less than in the MSR or pari-mutuel, though it is difficult to interpret this observation because *all* Bayes-inconsistent outcomes lead to an inference failure, regardless of their magnitude.

Conditional on observing a Bayes-inconsistent output, the poll and pari-mutuel are more likely to generate inconsistencies with $h_T(H) > 0.4$ than with $h_T(H) < 0.2$; all 6 of the pari-mutuel's Bayes-inconsistencies and 8 of the poll's 11 Bayes-inconsistencies have $h_T(H) > 0.4$. The double auction and MSR split the two types of errors evenly, with three of five periods giving $h_T(H) > 0.4$ for the double auction and four of seven giving $h_T(H) > 0.4$ for the MSR. Thus, the pari-mutuel and poll are somewhat handicapped by a tendency toward a uniform distribution, as would be predicted by the well-documented favorite-longshot bias (see, e.g., Ali 1977).[28]

4.3.5. Summary. In three of our four measures (error, no trade, and Bayes-inconsistencies), we found one mechanism to be uniquely bad and the others to be roughly equivalent. Specifically, the MSR generates the most error, the pari-mutuel generates the most no-trade periods, and the poll is the most frequently Bayes-inconsistent. The four mechanisms are roughly equal in the frequency with which mirages occur. The only mechanism that performed well in all measures (or, did not perform poorly in any one measure) is the double auction mechanism. A summary of the results appears in the second through fifth columns of Table 11.

[26] In several periods we do observe "meaningless" trade where a trader submits a wager in the final second before the market closes. If an individual is the only trader to place a wager in a pari-mutuel mechanism and does so at the last second, he faces no risk as long as he owns at least one of each security because he is effectively betting against himself. Thus, these trades are not informative (nor financially consequential) and are discarded from the analysis.

[27] We do find that, across all mechanisms, Bayes-inconsistent outcomes are significantly more likely to occur in the first period. No other period effects have been observed.

[28] We thank an anonymous referee for suggesting we explore favorite-longshot biases in our data.

RIGHTSLINK

Downloaded from informs.org by [132.174.255.116] on 14 November 2013, at 06:45 . For personal use only, all rights reserved.

Table 6 *p*-Values of Mechanism-by-Mechanism Wilcoxon Tests on the Distance to the Full-Information Posterior for the Eight-State Experiments

Eight states	Avg. distance	Dbl. auction	Mkt. scoring rule	Pari-mutuel	Poll
Avg. distance	—	0.696	0.527	0.605	0.418
Dbl. auction	0.696	—	**0.002**	*0.093*	**<0.001**
Mkt. scoring rule	0.527	—	—	*0.083*	0.324
Pari-mutuel	0.605	—	—	—	**0.001**
Poll	0.418	—	—	—	—

Notes. 10% Significance ordering: DblAuc $\succ$ Pari $\succ$ MSR $\succeq$ Poll. Italicized (boldfaced) entries are significant at the 10% (5%) level.

4.4. The Complex Environment: Eight States

4.4.1. Mechanism Accuracy. As with the two-state experiments, we measure a mechanism's error as the Euclidean (l_2) distance between the mechanism output distribution and the full-information posterior. The distribution of errors for each mechanism is compared against that of each other mechanism using a Wilcoxon rank sum test. This pairwise comparison procedure generates a significance ordering that ranks the mechanisms by their average errors.[29] The result of this procedure is reported in Table 6. The accuracy results for the eight-state experiments can be summarized by the significance statement "DblAuc $\succ$ Pari $\succ$ MSR $\succeq$ Poll," which indicates that the double auction is uniquely the worst mechanism (according to this error measure), the pari-mutuel is uniquely the second-worst, and the MSR and poll generate the lowest errors on average, with no significant difference between them.

As in the two-state environment, these results may be biased by the presence of cohort effects. Wilcoxon tests on each pair of sessions in each mechanism find 8 of 24 session pairs with significant differences in error at the 10% level (3 each in the double auction and MSR and 1 each in the pari-mutuel and poll). To account for these cohort effects we take a very conservative approach and view the average error distance of each session as a single observation, reducing our sample size to only four observations per mechanism.[30] Despite this dramatic loss in testing power, we still achieve two significant results: the poll's average error is significantly lower than both the double auction (p-value: 0.0286) and the pari-mutuel (p-value: 0.0571). This occurs because the highest error of the four poll sessions is still lower than the lowest error of the four double auction sessions and the lowest error of the four pari-mutuel sessions.[31] No other comparisons of session-level errors are significant at the 10% level.

Controlling for between-cohort variability using a nested ANOVA analysis alters the significance results slightly; the significance statements from that analysis are "DA $\succeq$ Pari $\succeq$ MSR $\succ$ Poll" and "DA $\succ$ MSR," and all significant results are significant at the 5% level. Thus, the poll is significantly better than all three competing mechanisms, and the double auction is significantly worse than all but the pari-mutuel.

4.4.2. Catastrophes: No Trade. In the eight-state experiments, no-trade periods were observed only in the pari-mutuel mechanism. One group of subjects traded in none of the eight periods and another group failed to trade in their fifth period. As with the two-state data, panel data techniques are unable to reliably disentangle mechanism effects from session effects because nearly all incidences of no-trade occur in a single session. The qualitative evidence, however, is sufficiently suggestive to lead us to conclude that the pari-mutuel mechanism is more susceptible to no-trade than the other three mechanisms. This conclusion is easily verified by binomial tests that incorrectly assume independence across all periods.

4.4.3. Catastrophes: Mirages. Recall that we define a mirage to be a mechanism output distribution that lies in an opposite direction from the prior as the full-information posterior. Mathematically, this occurs when $(h - p^0) \cdot (p^{FI} - p^0) < 0$; this is demonstrated in panel (B) of Figure 1.

Looking at the frequency of mirages (see Table 7), the double auction is most prone to mirage outcomes whereas the poll is the least prone. In a probit random effects test, the double auction is significantly worse than the poll (p-value: 0.009) but insignificantly worse than the other two mechanisms.

Comparing the angles between the vectors $(h - p^0)$ and $(p^{FI} - p^0)$ and applying pairwise Wilcoxon tests (see Table 8), we see that the double auction is uniquely the worst mechanism by this measure because its average output distribution points in a direction farthest from the full-information posterior. In fact, its average angle is nearly 90 degrees, indicating that the mechanism provides little to no information that is not already contained in the prior. In contrast, the other mechanisms do, on average, move toward the full-information posterior, indicating that

[29] In contrast to the results in the simple environment and based on a simple counting measure, we do find some evidence that prices are biased in favor of long-shots in the complex environment. This holds for all mechanisms, but is strongest for the double auction and poll. We believe that this finding likely confounds a number of sources of error, and we do not claim that we are able to identify this as a major cause of the poor performance of the mechanisms.

[30] The pari-mutuel has one session with no trade, leaving only three session-level observations.

[31] The difference in p-values between these two comparisons stems only from the fact that the pari-mutuel has one entire session with no trade and therefore only *three* session-level observations available.

Downloaded from informs.org by [132.174.255.116] on 14 November 2013, at 06:45 . For personal use only, all rights reserved.

Table 7 Number of Periods in Each Session (Out of 8) and Number of Periods Total (Out of 32) in Which Each Type of Catastrophic Failure Occurs in the Eight-State Experiments

	Dbl. auction		Mkt. scoring rule		Pari-mutuel		Poll	
	(S5, S6, S7, S8)	Tot.	(S3, S4, S1, S2)	Tot.	(S1, S2, S3, S4)	Tot.	(S7, S8, S5, S6)	Tot.
No trade	(0, 0, 0, 0)	0	(0, 0, 0, 0)	0	(0, 0, 8, 1)	9	(0, 0, 0, 0)	0
Mirage	(3, 1, 4, 4)	12	(1, 1, 2, 3)	7	(3, 1, 0, 3)	7	(0, 1, 2, 0)	3
None	(5, 7, 4, 4)	20	(7, 7, 6, 5)	25	(5, 7, 0, 4)	16	(8, 7, 6, 8)	29

Note. Every mechanism is Bayes-inconsistent in every period.

all mechanisms other than the double auction do provide more information than the prior, or the prior plus random noise.

A third way to measure the incidence of mirages is simply to count the number of dimensions of $(h - p^0)$ that have the same sign as the corresponding dimension of $(p^{FI} - p^0)$, excluding the first and last dimension since, in theory, they should not change. Table 9 reports the p-values of the pairwise Wilcoxon tests on the number of dimensions. The results are in line with the other two measures; the double auction is uniquely the most prone to mirages and the other three mechanisms do not significantly differ in the frequency or magnitude of observed mirages.

4.4.4. Catastrophes: Bayes-Inconsistency. Recall that an output distribution is labeled "Bayes-inconsistent" if it does not lie in the convex hull of the limit posteriors. In the eight-state case, distributions live in $\mathbb{R}^8$, but because the first and last dimensions should never differ from the prior, the convex hull lives in the six-dimensional subspace where those two dimensions are fixed at the prior level. Thus, an output distribution is automatically "Bayes-inconsistent" if either the first or last dimension differs from the prior. See Figure 2 for a simplified representation of this issue. In practice, Bayes-inconsistency occurs in every period under every mechanism in our eight-state experiments precisely because these first and last dimensions never perfectly match the prior probabilities, therefore indicating Bayes-inconsistency with a binary indicator variable is not informative. Therefore, we measure the "degree" of inconsistency as the distance between the output distribution and the convex hull. Using pairwise Wilcoxon tests (see Table 10), we find that neither the MSR nor the poll have significantly greater median distances than any other mechanism and that the double auction and pari-mutuel do have significantly greater median distances than at least one other mechanism. Thus, the MSR and the poll are less prone to large deviations from the convex hull.

An alternative way to measure the propensity for Bayes-inconsistency is to count the number of periods in which the distance between the output distribution and the convex hull is within ϵ for each ϵ greater than zero. The resulting graph of frequencies versus ϵ for each mechanism appears in Figure 5. The MSR and the poll generate output distributions within ϵ of the convex hull most frequently when ϵ is small. As

Table 8 p-Values of Mechanism-by-Mechanism Wilcoxon Tests Comparing the Angle (in Degrees) Between the Mechanism Output $(h - p^0)$ and the Full-Information Posterior $(p^{FI} - p^0)$

Eight states	Avg. angle	Dbl. auction	Mkt. scoring rule	Pari-mutuel	Poll
Average angle	—	89.23	66.12	74.68	69.07
Dbl. auction	89.23	—	**<0.001**	**0.011**	**<0.001**
Mkt. scoring rule	66.12	—	—	0.180	0.773
Pari-mutuel	74.68	—	—	—	0.286
Poll	69.07	—	—	—	—

Notes. 10% Significance ordering: MSR $\succeq$ Poll $\succeq$ Pari $\succ$ DblAuc. Larger values imply more error.

Table 9 p-Values of Mechanism-by-Mechanism Wilcoxon Tests Comparing the Number of Dimensions (Out of 6) of the Mechanism Output That Move in the Same Direction (from the Prior) as the Full-Information Posterior

Eight states	Avg. no.	Dbl. auction	Mkt. scoring rule	Pari-mutuel	Poll
Avg. no. dim.	—	2.69	3.69	3.70	3.97
Dbl. auction	2.69	—	**0.002**	**0.003**	**<0.001**
Mkt. scoring rule	3.69	—	—	0.798	0.239
Pari-mutuel	3.70	—	—	—	0.467
Poll	3.97	—	—	—	—

Note. 10% Significance ordering: Poll $\succeq$ Pari $\succeq$ MSR $\succ$ DblAuc.

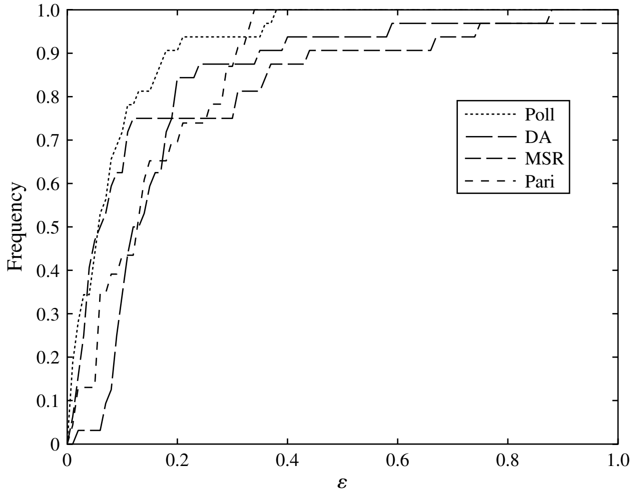
Table 10 p-Values of Mechanism-by-Mechanism Wilcoxon Tests Comparing the Severity of Bayes-Inconsistency, as Measured by the Distance Between the Mechanism Output Distribution and the Convex Hull of the Limit Posteriors

Eight states	Avg. dist	Dbl. auction	Mkt. scoring rule	Pari-mutuel	Poll
Avg. distance	—	0.447	0.362	0.398	0.312
Dbl. auction	0.447	—	**0.001**	0.107	**< 0.001**
Mkt. scoring rule	0.362	—	—	0.180	0.257
Pari-mutuel	0.398	—	—	—	**0.008**
Poll	0.312	—	—	—	—

Note. 10% Significance ordering: DblAuc $\succeq$ Pari $\succeq$ MSR $\succeq$ Poll, DblAuc $\succ$ MSR $\succeq$ Poll, DblAuc $\succeq$ Pari $\succ$ Poll.

Downloaded from informs.org by [132.174.255.116] on 14 November 2013, at 06:45 . For personal use only, all rights reserved.

Figure 5 Frequency of Periods (with Trade) in Which Bayes-Inconsistency Is Less Than ϵ



ϵ is increased, however, the MSR moves from most frequent to least frequent and the pari-mutuel moves from second-least frequent to most frequent. In other words, the MSR output tends to lie either very close to the convex hull or very far, whereas the pari-mutuel output consistently lies an intermediate distance from the convex hull. Thus, a market observer who is concerned about extreme levels of Bayes-inconsistency should prefer the pari-mutuel mechanism over the MSR in the eight-state environment. As for the double auction mechanism, however, the results are poor in either measure; its average distance from the convex hull is the highest, and the frequency with which it lands within ϵ of the convex hull is typically the lowest or second-lowest among the four mechanisms.

4.4.5. Summary. As with the two state case, we found one or two mechanisms to be uniquely bad according to each of our four measures (error, no trade, mirages, and Bayes-inconsistency), though the poorly performing mechanism varies with the measure. Specifically, the double auction and pari-mutuel generate larger errors, the pari-mutuel is the most prone to no trade, the double auction creates the most mirages, and the double auction and pari-mutuel generate the greatest amount of Bayes-inconsistency. The two mechanisms that did not perform poorly in any of the four measures are the poll and the MSR. Between these two the poll appears to outperform the MSR, though at statistically insignificant levels. The results for the eight-state experiments are summarized in the sixth through ninth columns of Table 11.

5. Five Observations

The results indicate that the poll and (to a somewhat lesser extent) the MSR perform well and the double auction poorly in the more complex environment. This raises the deeper question of *why* this occurs. What features of the poll and MSR make them successful that are not shared by the double auction? Based on our analysis of the data we state five observations about these three mechanisms that we believe are primarily responsible for the performance differences.

Observation 1. *Preferences are aligned in the poll, so traders have no incentive to misrepresent their information, whereas truth-telling in the MSR is weakly incentive compatible.*

A subject misrepresents his private information when he takes an action intended to send a false signal of his private information. Misrepresentation can interfere with the performance of a mechanism by adding noise to the public signals sent by a subjects actions. Although the potential for misrepresentation in equilibrium is a difficult question, it is clear that misrepresentation might present profit opportunities in mechanisms where incentives are not aligned. In the poll, however, a subject's payoff will generally increase in the quality of the information available to other subjects. Thus, the poll may be less subject to problems with misrepresentation. In the MSR a subject may have an incentive to misrepresent early, but his final announcement (if he believes it to be his final announcement) should be truth-telling; see the discussion of the MSR at the end of §2.

We construct a rough measure of misrepresentation as follows. Recall that each mechanism generates a

Table 11 Summary of Results

	Two states				Eight states			
Summary	Error	No trade	Mirage	Inconsistent	Error	No trade	Mirage	Inconsistent
Dbl. auction	✓	✓	✓	✓	×	✓	×	×
MSR	×*	✓	✓	✓	✓	✓	✓	✓
Pari-mutuel	✓	×*	✓	✓	×	×*	✓	×
Poll	✓	✓	✓	×*	✓	✓	✓	✓

Notes. A ✓ indicates the mechanism was not significantly outperformed by some other mechanism in that measure and an × indicates that it was. An ×* denotes either marginal significance (all p-values less than but close to 0.10) or cases where proper statistical tests were unavailable.

Downloaded from informs.org by [132.174.255.116] on 14 November 2013, at 06:45 . For personal use only, all rights reserved.

Table 12 Number of Observations of Misrepresentation by Mechanism

Mechanism	No. of misrepresentations
Dbl. auction	14
Mkt. scoring rule	5
Pari-mutuel	12
Poll	3

sequence of distributions $\{h_t\}_{t=0}^{T}$. An action at time t is said to move the posterior toward the full-information posterior if $\|h_t - p^F\| \leq \|h_{t-1} - p^F\|$; otherwise, the action moves the posterior away.[32] A subject is identified as a misrepresenter in a period if his moves include at least one move toward the full-information posterior and at least one move away, and all moves away precede all moves toward. We have 96 opportunities to observe misrepresentation for each mechanism (three subjects each in a total of 32 periods). The number of misrepresentations in each mechanism are presented in Table 12. We observe the fewest instances of misrepresentation in the poll and MSR. This is consistent both with the aligned incentives in the poll and the weak incentive compatibility built into the MSR.

Observation 2. *Traders have an incentive to participate in both the poll and the MSR because they are subsidized.*

Excluding the value of initial endowments, the double auction is a zero-sum game. A trader who does not participate earns her expected value of participating and, according to the no-trade theorem argument, she strictly prefers nonparticipation if rationality is common knowledge and she is risk averse. Even if rationality is not known, only those traders who expect to perform better than average will prefer to participate. Although trade occurs in every period in our data, there are 4 periods (of 64) where one of the three traders abstains from trading.

In the poll, however, there is no benefit to abstention; any trader can (weakly) improve the group's average report (relative to his posterior beliefs) by appropriately incorporating his private information into his own final report. Improving the final average report improves the payoff of everyone in the group.[33] Similarly, the MSR involves a subsidy when participants perform well as a group.

Observation 3. *Traders in the poll must submit entire probability distributions, preventing them from focusing on a small number of securities.*

It appears that market thinness in the eight-state world prevents the double auction from aggregating information properly. We find that there are only 2.60 transactions per minute across all markets in the eight-state environment, compared to 5.00 transactions per minute in the two-state environment; traders are trading half as frequently in the eight-state environment despite the fact that there are four times as many markets. Interestingly, total volume per minute is much higher in the eight-state environment (14.47 units per minute compared to 6.48 units per minute in the two-state environment), indicating that traders in the eight-state environment are making a small number of large transactions. Trades in the eight-state environment tend to focus on a small number of securities. Averaging across the four double auction sessions, trade on the two most active securities accounted for 46% of the transactions while trade on the two least active securities accounted for only 8% of the transactions.[34]

We conjecture that subjects focusing on a small number of securities indicates that attention is a constraint that binds in mechanisms that require separate focus on each event or security. In the double auction, subjects must analyze the market for each security separately. Given bounded attention, subjects are likely to focus or coordinate on a small number of securities, forgoing profits on other securities. Thus, we should expect some market prices to be far from equilibrated. To examine this conjecture, we consider the states *TTT* and *HHH*, whose posterior probabilities equal the prior probabilities of 24/75 and 4/75, respectively, because the ordering of the coins obviously does not affect the probability of these two states. If market prices are far from these values then profit opportunities may exist in these markets. In fact, we observe that the average distance between the final price and the prior probability is 13% for *TTT* and 7.6% for *HHH*. Both of these distances are significantly greater than the distances for any other mechanism (Wilcoxon p-values of <0.001).

Observation 4. *The poll averages the elicited beliefs, so the effects of a single aberrant trader are mitigated.*

Our final observation is that the poll performs relatively well compared to the other mechanisms because of lessened sensitivity to erroneous last actions. To identify the frequency of large errors in individual reports, we first calculate the average error in final

[32] For the poll, actions are ordinal, and we adopt the convention that $t \in \{0, 2, 4, 6, 8\}$ represent individual reports and $t \in \{1, 3, 5, 7, 9\}$ represent aggregate reports. Therefore, h_t is not unique when t is even. The different timing structure for the poll makes formal statistical comparisons difficult.

[33] The average payoff per trader per period in the poll is 25.9 cents and 35.0 cents for the two-state and eight-state treatments, respectively.

[34] There does not appear to be a systematic trend in which securities were traded the most or least frequently.

Downloaded from informs.org by [132.174.255.116] on 14 November 2013, at 06:45 . For personal use only, all rights reserved.

Table 13 **Number of Periods with Far-Off Last Report and Final Prediction**

	Two states		Eight states	
Mechanism	Last report	Output distribution	Last report	Output distribution
Dbl. auction	11	11	24	24
Mkt. scoring rule	18	18	9	9
Pari-mutuel	11	11	9	9
Poll	28	8	21	8

predictions across all the mechanism. Using the normalized l_ρ, the size of average errors in the two-state experiments is 0.155; in the eight-state experiments it is 0.5996. We define a period with far-off last report as one where the last action implies an individual posterior with a larger-than average prediction error. Because the poll requires all three individuals to submit their report simultaneously, we use the report with the largest prediction error as the last report. The number of periods with far-off reports in each mechanism are presented in Table 13.

In the double auction, pari-mutuel, and MSR, this last report has a direct effect on the mechanism's output. The number of periods with a far-off prediction—defined as a prediction with larger-than-average error—is necessarily the same as the number of periods with far-off reports. However, the poll balances out this errant last report by averaging it with the other two players' reports.

Despite the large numbers of far-off last reports, the poll produces the fewest instances of far-off final predictions. This is consistent with our claim that averaging in the poll makes it less sensitive to individual errors at end of the period. Note that if the players' final reports are derived from the same distribution, Jensen's inequality and the convexity of our error measure will imply lower prediction error for the poll. Another interesting observation from Table 13 is that the number of far-off last reports in the poll is among the highest compared to other mechanisms, which point to the possibility of players strategically using the averaging mechanism to offset expected error in other players' reports.

The risk that a far-off last report can unduly influence the outcome of the continuous mechanisms suggests that smoothing over the actions near the end of the period might provide an improvement over focusing exclusively on the final outcome of the mechanism. To evaluate this possibility, we average over the outcomes implied by the final 20% and 50% of actions in each of these mechanisms. For both the double auction and the pari-mutuel, this has no appreciable effect on the performance of the mechanism, as measured by average distance. For the MSR, however, smoothing over the last 20% of moves in the simple environment and smoothing over the last 50% of moves in the complex environment bring substantial improvements in performance. In the simple environment, the average error of the MSR drops from 0.210 to 0.119, whereas in the complex case, the error drops from 0.527 to 0.411. In both of these cases, the alternative degree of smoothing provides no appreciable improvement in performance. Although statistical comparisons between these smoothed outputs and the original outputs is inappropriate because of concerns about data mining, we note that the smoothed MSR produces the smallest average error for the simple environment, though this error would not have been significantly different than the errors in any of the other mechanisms. In the complex environment, the errors for the smoothed MSR would have made that mechanism statistically indistinguishable from the poll in the sense that both the poll and the smoothed MSR would outperform the other mechanisms at the same significance level. We also note, however, that the optimal degree of smoothing seems to vary with the complexity of the environment, leaving no obvious recommendation for a better means to evaluate the output of the MSR. Regardless, we do believe these results suggest the possibility of designing a mechanism that exploits the weak incentive compatibility of the MSR while also generating an output that is more robust to the behavioral characteristics identified here.

Observation 5. *Transactions in the double auction are bilateral; in all other mechanisms transactions are executed unilaterally. The double auction is therefore more labor intensive.*

Because a transaction in the double auction requires the active involvement of two parties, it is simply a more labor-intensive mechanism. With a small number of traders whose time is either constrained or costly, it is reasonable to expect that information aggregation will be inhibited by the fact that subjects must seek out trading partners in every transaction. By contrast, the poll simply requires that each trader send a fixed number of discrete messages, and the market scoring rule and pari-mutuel effectively use market makers that allow traders to act without coordinating with other traders. An analysis of transaction times in the double auction reveals that traders may well have been time constrained; there is no perceptible change in transaction volumes toward the end of the five-minute periods. Given enough time, the mechanism's performance may significantly improve. In many field applications, however, labor cost and time constraints are very real issues that may hinder the double auction's ability to aggregate information and generate useful predictions.

RIGHTSLINK
Copyright Clearance Center

Downloaded from informs.org by [132.174.255.116] on 14 November 2013, at 06:45 . For personal use only, all rights reserved.

6. Discussion

In comparing these four mechanisms (the double auction, the market scoring rule, the pari-mutuel, and the poll), we find that the performance of the mechanisms is significantly affected by the complexity of the environment. In particular, the double auction mechanism appears to perform relatively better when the number of states is small relative to the number of traders and the inference problem of inverting beliefs back into received signals and then converting aggregated signals into an aggregated belief is relatively easy. When the environment becomes more complicated, both in the number of states and in the difficulty of the inference problem, the performance of the double auction market breaks down and other mechanisms emerge as superior. In particular, the iterative poll is the only mechanism in our experiment that was not outperformed by some other mechanism in any of the four measures of error considered.

Identifying which mechanisms perform well in given environments is only the first step in this research. The most compelling line of inquiry is into the underlying reasons for a mechanism to succeed or fail in a given environment. For example, we observe that the failure of the double auction in the eight-state experiments is due in part to the increased ratio of the number of securities to the number of traders: the "thin markets" problem. As the number of securities exceeds the number of traders, agents apparently focus their limited attention on a small subset of the securities during the trading period. This creates an additional coordination problem as traders seek to focus their attention on markets in which trading is currently most profitable, perhaps because of the trading volume in that market and the private information of the given trader. If some securities are ignored and receive no trades, then information aggregation is necessarily incomplete.

One open question is how these mechanisms would perform if the number of traders were increased beyond three. In previously unpublished pilot experiments, Joel Grus and John Ledyard (see Ledyard 2005) compare the same four mechanisms (double auction, MSR, pari-mutuel, and poll) in a two-state environment similar to ours using 3, 7, and 12 participants.[35] The Grus–Ledyard experiments do not include the eight-state design. Agents participate in the same group of n subjects for the entire experiment. Each group participates in three mechanisms for eight periods each, as opposed to two mechanisms per group in our design. Their measure of error uses the Kullback–Leibler information criterion ("KL distance") instead of the Euclidean distance (Kullback and Leibler 1951; this follows Ledyard et al. 2009), though in our data these two measures are highly correlated. The average KL distances are summarized in Table 14.

Table 14 **Average Errors (Using KL Distance) from the Grus–Ledyard Pilot Data**

No. of participants	3	7	12
Dbl. auction	0.243	0.198	0.016
MSR	0.045	0.001	0.000
Pari-mutuel	0.158	0.019	0.006
Poll	0.046	0.004	0.001

Unsurprisingly, each mechanism becomes more accurate as the number of traders increases.[36] The absolute improvement is larger for those mechanisms with larger errors (double auction and pari-mutuel), but the percentage improvement per additional trader is 10.4%, 11.1%, 10.7%, and 10.9% for the double auction, MSR, pari-mutuel, and poll, respectively.[37] This suggests that, for the simple two-state environment, increases in the number of traders will have little effect on the relative performance of these mechanisms. Whether this similarity extends to the more complex environment is an open question, though the success of fairly complex double auction prediction markets with many traders (such as TradeSports) suggests that the double auction eventually does benefit differentially from increased thickness.[38]

Many other open questions remain. Fine details such as the complexity of the information structure could be altered and results compared. New mechanisms or perturbed versions of these mechanisms could be compared in the laboratory. The context of various business environments could be overlaid on our sterile laboratory environment to explore particular real-world implementations. On the theoretical front, little is known about the equilibrium and manipulability of these mechanisms played by fully rational agents, let alone boundedly rational agents prone to various biases and cognitive errors.

[35] The major differences between their design and ours are the number of subjects, $f(X) = f(Y) = 1/2$, $f(H \mid X) = 0.2$, and $f(H \mid Y) = 0.8$, and subjects always see two sample flips for their private information. Their periods also lasted five minutes, though their poll with 12 participants ran through five iterations instead of three. The seven-subject sessions of the pari-mutuel and MSR actually had eight subjects.

[36] This improvement in accuracy is also correlated with increases in the number of trades per minute per subject.

[37] The Grus–Ledyard data indicates that in the two-state case the double auction gains less when the number of traders is still small and more as the number becomes larger. Data on the effect of increasing cohort size on the eight-state environment is unavailable.

[38] Grus and Ledyard also examined subsidized versions of the double auction and pari-mutuel. They find that subsidies significantly improve the accuracy of both mechanisms, supporting our Observation 2.

RIGHTSLINK

Downloaded from informs.org by [132.174.255.116] on 14 November 2013, at 06:45 . For personal use only, all rights reserved.

Our larger goal with this research is to help develop the practice of behavioral mechanism design, where behavioral insights inform both the design of mechanisms for the immediate future and the modification of theories that can be used to find optimal mechanisms for practical applications into the future.

Acknowledgments

This work was partially supported by National Science Foundation Grant SES-0847406. The authors thank Rachel Croson, Catherine Eckel, John Fountain, Rick Green, Joel Grus, Glenn Harrison, Ernan Haruvy, Shimon Kogan, Tony Kwasnica, Sherry Li, Bryan Routledge, Justin Wolfers, an associate editor, and two anonymous referees for their helpful comments and conversations. Ines Fiorelli and Andrej Svorencik provided useful research assistance.

References

Ali, M. M. 1977. Probability and utility estimates for racetrack bettors. *J. Political Econom.* **85**(4) 803–815.

Berg, J., R. Forsythe, F. Nelson, T. Rietz. 2008. Results from a dozen years of election futures markets research. C. R. Plott, V. L. Smith, eds. *Handbook of Experimental Economics Results*, Vol. 1. Elsevier B.V., Amsterdam, 742–751.

Bloomfield, R., M. O'Hara, G. Saar. 2009. How noise trading affects markets: An experimental analysis. *Rev. Financial Stud.* **22**(6) 2275–2306.

Camerer, C. F. 1998. Can asset markets be manipulated? A field experiment with racetrack betting. *J. Political Econom.* **106**(3) 457–481.

Camerer, C. F., K. Weigelt. 1991. Information mirages in experimental asset markets. *J. Bus.* **64**(4) 463–493.

Chen, K.-Y., L. R. Fine, B. A. Huberman. 2001. Forecasting uncertain events with small groups. *Proc. 3rd ACM Conf. Electronic Commerce*, ACM, New York, 58–64.

Chen, Y., D. M. Reeves, D. M. Pennock, R. D. Hanson, L. Fortnow, R. Gonen. 2007. Bluffing and strategic reticence in prediction markets. *Proc. 3rd Internat. Conf. Internet and Network Econom.*, Springer-Verlag, Berlin, 70–81.

Cowgill, B., J. Wolfers, E. Zitzewitz. 2009. Using prediction markets to track information flows: Evidence from Google. S. Das, M. Ostrovsky, D. Pennock, B. K. Szymanski, eds. *Auctions, Market Mechanisms and Their Applications*, Vol. 14. Springer, Berlin, 3.

DeLong, J. B., A. Shleifer, L. H. Summers, R. J. Waldmann. 1990. Positive feedback investment strategies and destabilizing rational speculation. *J. Finance* **45**(2) 379–395.

Forsythe, R., R. Lundholm. 1990. Information aggregation in an experimental market. *Econometrica* **58**(2) 309–347.

Fréchette, G. R. 2009. Laboratory experiments: Professionals versus students. Working paper, New York University, New York.

Gjerstad, S. 2005. Risk aversion, beliefs, and prediction market equilibrium. Working paper, University of Arizona, Tucson.

Hanson, R. 2003. Combinatorial information market design. *Inform. Systems Frontiers* **5**(1) 107–119.

Hanson, R. 2007. The policy analysis market: A thwarted experiment in the use of prediction markets for public policy. *Innovations* **2**(3) 73–88.

Hanson, R., R. Oprea. 2009. A manipulator can aid prediction market accuracy. *Economica* **76**(302) 304–314.

Hanson, R., R. Oprea, D. Porter. 2006. Information aggregation and manipulation in an experimental market. *J. Econom. Behav. Organ.* **60**(4) 449–459.

Hopman, J. W. 2007. Using forecasting markets to manage demand risk. *Intel Tech. J.* **11**(2) 127–136.

Kullback, S., R. A. Leibler. 1951. On information and sufficiency. *Ann. Math. Statist.* **22**(1) 79–86.

Kyle, A. S. 1985. Continuous auctions and insider trading. *Econometrica* **53**(6) 1315–1335.

Ledyard, J. O. 2005. Information markets. Nancy L. Schwartz Memorial Lecture, Kellogg Graduate School of Management, Northwestern University, Evanston, IL.

Ledyard, J. O., R. D. Hanson, T. Ishikida. 2009. An experimental test of combinatorial information markets. *J. Econom. Behav. Organ.* **69**(2) 182–189.

Manski, C. F. 2006. Interpreting the predictions of predictions markets. *Econom. Lett.* **91**(3) 425–429.

McKelvey, R. D., T. Page. 1990. Public and private information: An experimental study of information pooling. *Econometrica* **58**(6) 1321–1339.

Milgrom, P., N. Stokey. 1982. Information, trade and common knowledge. *J. Econom. Theory* **26**(1) 17–27.

Neter, J., M. H. Kutner, C. J. Nachtsheim, W. Wasserman. 1996. *Applied Linear Statistical Models*, 4th ed. McGraw-Hill, New York.

Nikolova, E., R. Sami. 2007. A strategic model for information markets. *Proc. 8th ACM Conf. Electronic Commerce*, ACM, New York, 316–325.

O'Brien, J., S. Srivastava. 1991. Dynamic stock markets with multiple assets: An experimental analysis. *J. Finance* **46**(5) 1811–1838.

Plott, C. R. 2000. Markets as information gathering tools. *Southern Econom. J.* **67**(1) 2–15.

Plott, C. R., K.-Y. Chen. 2002. Information aggregation mechanisms: Concept, design, and implementation for a sales forecasting problem. Caltech Social Science Working Paper 1131, California Institute of Technology, Pasadena.

Plott, C. R., S. Sunder. 1982. Efficiency of experimental security markets with insider information: An application of rational-expectations models. *J. Political Econom.* **90**(4) 663–698.

Plott, C. R., S. Sunder. 1988. Rational expectations and the aggregation of diverse information in laboratory security markets. *Econometrica* **56**(5) 1085–1118.

Plott, C. R., J. Wit, W. C. Yang. 2003. Parimutuel betting markets as information aggregation devices: Experimental results. *Econom. Theory* **22**(2) 311–351.

Rhode, P. W., K. S. Strumpf. 2004. Historical presidential betting markets. *J. Econom. Perspect.* **18**(2) 127–142.

Rosenbloom, E. S., W. Notz. 2006. Statistical tests of real-money versus play-money prediction markets. *Electronic Markets* **16**(1) 63–69.

Selten, R. 1998. Axiomatic characterization of the quadratic scoring rule. *Experiment. Econom.* **1**(1) 43–61.

Servan-Schreiber, E., J. Wolfers, D. M. Pennock, B. Galebach. 2004. Prediction markets: Does money matter? *Electronic Markets* **14**(3) 243–251.

Shmaya, E., L. Yariv. 2008. Foundations for Bayesian updating. Working paper, California Institute of Technology, Pasadena.

Tetlock, P. C. 2008. Does liquidity affect securities market efficiency? Working paper, University of Texas at Austin, Austin.

Thaler, R. H., W. T. Ziembda. 1988. Anomalies: Parimutuel betting markets: Racetracks and lotteries. *J. Econom. Perspect.* **2**(2) 161–174.

Wolfers, J., E. Zitzewitz. 2004. Prediction markets. *J. Econom. Perspect.* **18**(2) 107–126.

Wolfers, J., E. Zitzewitz. 2006. Interpreting prediction market prices as probabilities. NBER Working Paper 12200, National Bureau of Economic Research, Cambridge, MA.

RIGHTSLINK