Elicitation of expectations using Colonel Blotto

Peeters, Ronald; Wolk, Leonard

published in
Experimental Economics
2019

DOI (link to publisher)
10.1007/s10683-018-9596-x

citation for published version (APA)
Peeters, R., & Wolk, L. (2019). Elicitation of expectations using Colonel Blotto. *Experimental Economics*, *22*(1), 268-288. https://doi.org/10.1007/s10683-018-9596-x

Experimental Economics (2019) 22:268–288
https://doi.org/10.1007/s10683-018-9596-x

ORIGINAL PAPER

Elicitation of expectations using Colonel Blotto

Ronald Peeters[1] · **Leonard Wolk**[2]

Received: 12 September 2017 / Revised: 13 November 2018 / Accepted: 15 November 2018 /
Published online: 22 November 2018

Abstract

We develop a mechanism based on the Colonel Blotto game to elicit (subjective) expectations in a group-based manner. In this game, two players allocate resources over possible future events. A fixed prize is awarded based on the amounts the players allocate to the realized event. We consider two payoff variations: under the proportional-prize rule, the award is split proportionally to the resources that players allocate to the realized event; under the winner-takes-all rule, the full award is given to the player who allocate the most resources to the realized event. When probabilities by which events realize are common knowledge to the players, both games are Bayesian–Nash incentive compatible in the sense that (expected) equilibrium allocations perfectly reflect the true realization probabilities. By means of a laboratory experiment, we find that in a setting where realization probabilities are common knowledge the game with the proportional-prize rule (Prop) elicits better distributions compared to both the winner-takes-all variation (Win) and a benchmark mechanism based on an individual-based proper scoring rule (Ind). Without common knowledge of realization probabilities Prop is at least as good as Ind, showing that it is possible to use a game to elicit expectations in a similar fashion to using a proper scoring rule.

Keywords Colonel Blotto · Expectation elicitation · Experiment · Behavioral mechanism design

JEL Classification C72 · C92 · D83 · D84

We thank Matt Embrey, Glenn Harrison, Georgia Kosmopoulou, Carlos Lamarche, Josh Miller, Paulo Somaini, Martin Strobel, Alexander Westkamp, the editor, two anonymous referees, and the audiences at the CEREC Workshop in Economics (Brussels, 2015), the Conference on Auctions, Competition, Regulation, and Public Policy (Lancaster, 2015), the Behavioral and Experimental Economics Symposium (Maastricht, 2015), the Experimental Finance Conference (Tucson, 2016; Heidelberg, 2018), the Australasian meeting of the Econometric Society (Auckland, 2018), VU Amsterdam as well as UMass-Lowell for very helpful comments and suggestions.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10683-018-9596-x) contains supplementary material, which is available to authorized users.

Extended author information available on the last page of the article

1 Introduction

Eliciting expectations about the future and constructing forecasts of future events are integral to successful business planning. In cases where for example employees possess information about the future that are not (yet) reflected in official performance metrics, it is integral that we are able to elicit and aggregate such information accurately. Recently, several market based mechanisms have been proposed; for instance, Baillon (2017) introduces Bayesian markets to elicit subjective beliefs and Gillen et al. (2017) show that forecasts constructed using a competitive forecasting mechanism outperform official sales forecasts at Intel.

In this study, we propose a strategic mechanism that resembles a betting pool for a single event and test its elicitation performance using data gathered in a laboratory experiment. The strategic component offers an environment with simple rules and where payoffs are determined transparently. Further, it relies on a small number of forecasters, which offers the possibility for the mechanism to be implemented in small groups, or teams, akin to what is typically observed in a business setting where relevant information may not be widely dispersed throughout an organization.

More precisely, the elicitation mechanism we propose concerns a variation of the Colonel Blotto game, where two players divide a given amount of resources over a set of possible future events and where final payoffs are determined by the resources allocated to the realized event. We implement two different payment rules, a winner-takes-all rule (Win), where the prize is awarded to the player(s) with the most resources on the realized event, and a proportional-prize rule (Prop), where the prize is shared in proportion to the resources allocated to the realized event. We show that, when probabilities by which events realize are common knowledge to the players, the games designed are Bayesian–Nash incentive compatible. Under the proportional-prize rule, there is a unique equilibrium in which both players allocate their resources in proportion to these realization probabilities. Under the winner-takes-all rule, in equilibrium, both players randomize their allocations in a way that the expected allocation of resources is proportional to the true realization probabilities. Hence, both variations of the Blotto game produce equilibrium properties that are appealing for elicitation and prediction practices.

While in practical applications the strategic uncertainty present in our games may negatively affect the mechanisms' performance, there are other factors embodied by strategic environments that have the potential to improve the performance of the mechanism. First, the game is easy to explain and to implement, which is important given that, in scope of endogenous participation, individuals are known to be attracted to simpler mechanisms (Carpenter et al. 2008). Second, the joy of winning in a game may trigger more cognitive effort, relative to a situation where forecasters are individually incentivized (for instance, by a proper scoring rule) or via a market mechanism (as in prediction markets).[1] Third, competition may incentivize players

[1] In this regard, research in educational psychology has shown that competitive environments increase performance (Ames 1984). Moreover, there is ample of evidence from the experimental economics literature that individuals perform significantly better in competitive environments (cf. Niederle and Vesterlund 2007).

to allocate resources to extreme events, in accordance with realization probabilities, that are otherwise easily underestimated. Other major advantages of this game, which are not unique to our proposed game, are that few individuals are required to produce accurate forecasts and that individual expectations are elicited as a density over the entire distribution of possible future events, rather than just as a mean or a median. The latter property is particularly interesting when there are a small number of participants, or when the density is not symmetric or unimodal.

Using data gathered in a laboratory experiment, we test the performance of the mechanisms using a third non-strategic mechanism (Ind) as a benchmark. In this treatment individuals are incentivized via a quadratic scoring rule (Brier 1950). Next to the three mechanisms, as one treatment dimension, we have a second informational treatment dimension. In the first informational variation, following theoretical predictions, the true realization probabilities are common knowledge to the players (Baseline). In the second variation, more relevant in view of practical applications, players gradually learn these probabilities via observations (Predict).

We find that the Blotto game augmented with the proportional-prize payment rule outperforms both other mechanisms in the Baseline information variation and its superiority relative to the Blotto game with winner-takes-all payoffs is not driven solely by randomizing behavior. In the Predict information variation, we find that the performances of the different mechanisms are not statistically distinguishable. This is striking, since strategic uncertainty about opponent behavior in the game potentially could distort the expectations that we elicit. When eliciting expectations using a proper scoring rule there is no such distortion and comparing proportional incentives to the proper scoring rule, neither of the two mechanisms appear to be worse than the other from any perspective.

Our paper closely relates to several streams of the literature. First, it relates to the elicitation of subjective information using incentive compatible mechanisms such as scoring rules (Brier 1950; Prelec 2004) as well as prediction markets (Forsythe et al. 1992). Scoring rules have been applied in a wide variety of fields and have shown their success in extracting subjective information (for an overview see Carvalho 2016). One advantage of a scoring rule is that it elicits beliefs on an individual level and it is thus free from any strategic concerns. We use such a mechanism as a benchmark for the strategic mechanisms in our experiments. Yet, market based mechanisms such as prediction markets have also shown to be successful in aggregating dispersed beliefs about a future event, and have been applied in a wide variety of business environments such as Google (Cowgill et al. 2009) and Hewlett-Packard (Chen and Plott 2002) and for scientific reproducibility (Dreber et al. 2015).

Second, our study relates closely to the literature on parimutuel betting (Figlewski 1979; Thaler and Ziemba 1988). In a parimutuel betting market, a bookmaker offers prices for future events that are set by the relative demand and/or odds of these events taking place (Plott et al. 2003). Within this stream of the literature, our paper most closely links to that of Gillen et al. (2017), who design a distributional forecasting mechanism and conduct a field test at Intel, a large semiconductor firm. The authors' mechanism closely resembles a parimutuel betting market where forecasters purchase tickets that can be spent on possible future outcomes. An interesting feature of their implementation is that the price of a ticket is not fixed but instead depends on the timing of the purchase. By inducing a cost of delay, this mechanism helps mitigate strategic behavior where betting takes place close to the end of the market. This is different from our study, which does not involve a time-dimension within each round and hence, in our experiment, there are no timing issues related to the placing of a 'bet'. The authors report strong results and show that the mechanism consistently outperforms official Intel forecasts, especially at short horizons. However, since it is implemented in the field, the source of the performance improvement is not fully clear. It could either be that the mechanism is able to aggregate information more efficiently or that it is able to collect superior information compared to official Intel forecasts. Our study complements that of Gillen et al. (2017), as we focus on the incentive structure that affects the revelation of expectations and, in turn, also the forecasting performance. Hence, our interest lies primarily in the design and performance of payment rules and not in the actual information gathering process, which takes place outside the game in our study.

Third, there is a large existing stream of literature on Blotto experiments, that includes four contributions that closely relate to ours.[2] Avrahami and Kareev (2009) conduct Blotto experiments investigating the role of asymmetries in players' strengths on their allocation decisions. Their symmetric benchmark treatments are in essence identical to our baseline treatment with a winner-takes-all payment rule, but with uniform realization probabilities. They find that players' behavior approximates the game-theoretic solution quite well (in particular for the benchmark treatment with equal player strengths). The experiments by Avrahami et al. (2014) also included treatments with non-uniform realization probabilities, and produced the finding that players' resource allocations correlate with the realization probabilities. Chowdhury et al. (2013) also consider a setting with asymmetries in players' strengths, both for the winner-takes-all payment rule and the lottery payment rule that is theoretically equivalent to our proportional-prize rule. In their set-up, players can collect a reward for each battlefield, which is theoretically equivalent to a setting like ours with uniform realization probabilities. They find that players' allocations of resources are in accordance with the theoretical predictions for both payment rules. Finally, the study by Duffy and Matros (2017) include treatments with symmetric players and a lottery payment rule that is theoretically equivalent to our proportional-prize rule. Similar to the study of Chowdhury et al. (2013), players can collect rewards for each battlefield, but the valuations are unequal across battlefields. They find mean allocations to be close to equilibrium predictions. Moreover, like in Avrahami et al. (2014), players' resource allocations correlate with the battlefield valuations.

The remainder of the paper is organized as follows. In Sect. 2, we present the two game variations and their equilibrium properties. Next, we present our experimental design and results in Sects. 3 and 4, relegating a further discussion on these results to Sect. 5. Section 6 concludes.

[2] For a review of the literature not addressed here, we refer to the studies that we do address.

2 Blotto game

There are two players, each of whom has a unit amount of resources that are to be distributed over $m \geq 3$ events $E_1, \dots, E_m$ with respective realization probabilities $p_1, \dots, p_m$, where $p_j \in (0, \frac{1}{2})$ for all $j = 1, \dots, m$ and $\sum_{j=1}^m p_j = 1$. We assume that the p_j probabilities are common knowledge to the players. The resources available to the players are of the use-it-or-lose-it kind, such that there is no benefit to the players for not using all their available resources. Let the distribution of player i be denoted by $\sigma^i = (\sigma^i_1, \dots, \sigma^i_m)$ with $\sigma^i_j \geq 0$ for all $j = 1, \dots, m$ and $\sum_{j=1}^m \sigma^i_j = 1$. There is a reward V that is allocated to the players depending on how both players have allocated their resources to the event that realizes.

We consider two possible payoff structures: a proportional-prize rule and a winner-takes-all rule. Under the proportional-prize rule, the reward is split over the players in proportion to the resources that they have allocated to the realized event. That is, player i receives a share of $\frac{\sigma^i_\ell}{\sigma^i_\ell + \sigma^{-i}_\ell}$ of the reward in case event E_ℓ realizes (which is the case with probability p_ℓ). Under the winner-takes-all rule, the full reward is assigned to the player that allocated the most resources to the realized event. The reward is evenly split among the players in case both have allocated the same (positive) amount of resources to this event. In either case, players receive nothing if none of the players has put any mass on the realized event.

Proposition 1 (Proportional-prize rule) There is a unique Nash equilibrium in which both players distribute their resources in proportion to the true realization probabilities; that is, $\sigma^i_j = p_j$ for all events $j = 1, \dots, m$, for both players $i = 1, 2$.

Proof First, notice that in equilibrium there does not exist an event j for which $\sigma^1_j = \sigma^2_j = 0$, since both players would benefit from shifting part of the mass that is put on any other event to this event j. Second, there does not exist a player i and an event j for which $\sigma^i_j = 0$. Suppose to the contrary that there does exist such a player and event. Then, by our first observation, we know that the other player must have strictly positive mass on this event: $\sigma^{-i}_j > 0$. Now, this other player benefits from shifting part of the mass that he put on event j to any other event on which player i has put positive mass (which must exist). This contradicts σ being an equilibrium. We conclude that in equilibrium both players put positive mass on all events.

In a Nash equilibrium $\bar{\sigma}$, player i's resource distribution $\bar{\sigma}^i$ solves

$$\max_{(\sigma^i_j)_{j=1,\dots,m}} \sum_{j=1}^m p_j \frac{\sigma^i_j}{\sigma^i_j + \bar{\sigma}^{-i}_j}$$

$$\text{subject to} \quad \sigma^i_j \geq 0 \;\; (j = 1, \dots, m) \quad \text{and} \quad \sum_{j=1}^m \sigma^i_j = 1$$

Since both players allocate positive amounts to all possible events, the first order conditions are given by $p_j\,\bar{\sigma}_j^{-i} = \lambda^i\,[\,\bar{\sigma}_j^i + \bar{\sigma}_j^{-i}\,]^2$ for all $j = 1, \dots, m$ and $i = 1, 2$. From this it follows that $\lambda^1\,\bar{\sigma}_j^1 = \lambda^2\,\bar{\sigma}_j^2$ for all $j = 1, \dots, m$. By summing these equalities, we find that $\lambda^1 = \lambda^2$, which implies that $\bar{\sigma}_j^1 = \bar{\sigma}_j^2$ for all $j = 1, \dots, m$. Exploiting this symmetry, the first order conditions can be rewritten as $p_j = 4\,\lambda^i\,\bar{\sigma}_j^i$ for all $j = 1, \dots, m$ and $i = 1, 2$. Summing these equations over $j = 1, \dots, m$ gives $\lambda^i = \frac{1}{4}$, leading to the conclusions that $\bar{\sigma}_j^i = p_j$ for all $j = 1, \dots, m$ and $i = 1, 2$.[3] □

According to the proposition, the Blotto game augmented with a proportional-prize rule is Bayesian–Nash incentive-compatible. This implies that, in a situation where an entity does not know the realization probabilities, but knows that they are common knowledge among some individuals, equilibrium behavior of these individuals when playing this game would perfectly reveal these probabilities. In this respect, the game has the potential to act as a proper elicitation mechanism for this entity.

Proposition 2 (Winner-takes-all rule) There does not exist an equilibrium in pure strategies and all randomizations over allocations with the property that for all events $j = 1, \dots, m$ the univariate marginal distributions are uniform on $[0, 2\,p_j]$ constitute a symmetric Nash equilibrium.

Proof Using the classification of Kovenock and Roberson (2015), our set-up is equivalent to the Colonel Blotto game in which battlefield valuations are heterogeneous across battlefields and symmetric across players and where players face symmetric resource constraints. This situation has been investigated by Gross (1950), Friedman (1958), Laslier (2002) and Thomas (2018) and we refer to Proposition 1 in Thomas (2018) for this result. □

According to this proposition, in the Blotto game augmented with a winner-takes-all rule, players can be expected to randomize over the possible distributions of their resources in a way that the expected amount of resources on each event is in proportion to the realization probability of this event; that is, $\mathbb{E}(\sigma_j^i) = p_j$ for all events $j = 1, \dots, m$, for both players $i = 1, 2$. Similar to the game with proportional-prize rule, this game also has elicitation potential. However, the property is not as strong here, since players have to resort to randomized strategies, such that ex-post realizations may not reflect the true realization probabilities accurately. The obvious way to improve on elicitation accuracy in practice is to have this game played by multiple pairs and aggregating their distributions.[4]

[3] The strategy profile in the theorem has already been shown to be a Nash equilibrium in Friedman (1958). The simpler proof presented here follows Robson (2005).

[4] To give an impression of the impact of aggregation over multiple pairs, Fig. A.1 in the online appendix shows cumulative distributions of the (numerically estimated) accuracy of the elicited distributions, measured by the Hellinger distance (as explained in Sect. 4), when aggregating over 32, 16, 8, 4, 2 and 1

1–10	11-20	21–30	31–40	41–50	51–60	61–70	71–80	81–90	91–100	101+
1	6	10	10	9	8	7	6	5	5	33

Fig. 1 The eleven events

3 Experiment

In this section, we present the 2×3 between-subjects factorial experimental design that we developed to assess the elicitation and prediction performance of the two game variations. Next to the two games, we implement an individual and incentive-compatible mechanism based on the quadratic scoring rule as a benchmark. Within each of the three payoff-based treatment variations we implement two informational conditions. One with common knowledge of the underlying realization probabilities and one where these probabilities are not commonly known. The first one closely follows theoretical assumptions with the aim to test the findings in the previous section. The second condition is less well theoretically founded but can instead be considered practically more relevant.

3.1 Setting

Individuals allocate 100 resources over eleven bins as they are illustrated in Fig. 1. These eleven bins represent the possible events (labels above the bins) of which precisely one will realize. A chance mechanism determines the bin that is decisive for the individuals' payoffs.

The chance mechanism that we implement is based on a random process that starts at a value of zero at time $t = 0$ and at each unit of time the value is incremented with a real number drawn randomly according to a normal distribution with mean zero and standard deviation equal to 0.2770. The process terminates either when the value crosses the lower boundary at -2.5, crosses the upper boundary at $+2.5$, or has reached time $t = 100$ without having reached one of these boundaries.[5] Figure 2 shows one time series generated by this process that led to a termination at the lower bound at time $t = 63$. The termination time uniquely determines the bin that is decisive for the payoffs. The labels above the first ten bins in Fig. 1 represent ranges of termination times; the eleventh bin represents the event that the process did not terminate before $t = 100$. So, in case of the time series in Fig. 2 the seventh bin is selected to be decisive for the payoffs.

For the process that we used in the experiment, Fig. 3 presents the true distribution over termination times, conditional on termination before $t = 100$. Moreover, the probability of the process not terminating before $t = 100$ equals one third.

Footnote 4 (continued)
pairs. The lower the Hellinger distance is, the better the accuracy. The figure shows that aggregation over pairs increases the accuracy at a decreasing rate.

[5] This random process is similar to the one used in Peeters and Wolk (2017) to study the interval scoring rule as a mechanism for individual expectation elicitation.

Fig. 2 An example of a time series

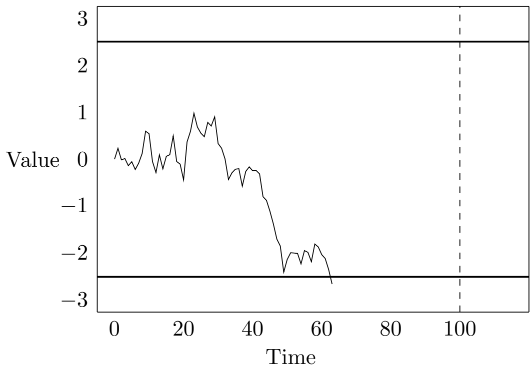
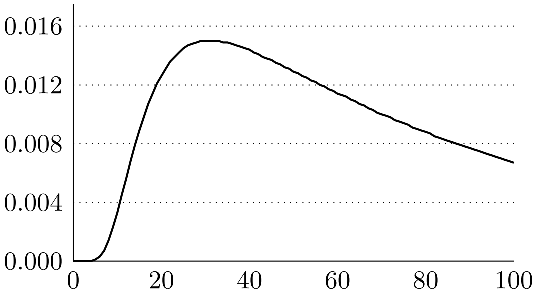


Fig. 3 Distribution over termination times conditional on termination before $t = 100$



The resource allocation displayed in the bins of Fig. 1 present the actual probability (expressed as percentages and rounded to the nearest integer) that the process will terminate at the respective bin.

3.2 Design

Individuals make their decisions repeatedly over the course of twenty rounds. Prior to the first round, participants were shown an animation of a randomly generated time series. After having seen this animation, they were asked to allocate their resources, in integer values, over the eleven bins. Next, they were shown the animation of the time series that was generated to select the decisive bin for the first round. Prior to the second round, participants received all payoff relevant information from the first round. This procedure continued until the last (twentieth) round.

Our first treatment variation concerns the incentive mechanism. In the first two variations, incentives are group-based. In these, subjects interact in pairs (pairs being fixed over the course of twenty decision periods) and the resources that the individuals have allocated to the decisive bin determine how the reward of 200 points is shared among them. In one of them (the Prop treatment), the shares of the reward that the individuals receive are proportional to the amount of resources that they allocated to the decisive bin. In the other (the Win treatment), the full reward is allocated to the individual that allocated the most resources to the decisive bin, with the reward being split evenly in case both individuals allocated the same amount of resources to this bin. In either case, individuals receive nothing when the total number of resources allocated to the decisive bin equals zero. In the third variation (the Ind treatment), incentives are individual-based, using the quadratic scoring rule (cf. Harrison et al. 2017). That is, they receive a payoff of $\alpha + \beta \cdot [\, 2\,\sigma_{\ell} - \sum_{j=1}^{11} \sigma_j^2 \,]$ points in case their allocation is according to the distribution $(\sigma_j)_{j=1}^{11}$ and bin ℓ is selected. In order to facilitate fair comparison across incentive mechanisms, we set α and β to 86.1623, such that in expectation 100 points are earned when individuals allocate their resources in accordance to the true realization probabilities and negative payoffs are avoided.[6]

In addition to the incentive mechanism as one treatment variation, we have the information condition as a second treatment variation. The Baseline treatment follows the theoretical model in Sect. 2 in that the true realization probabilities are common knowledge to the players. In the Predict treatment, in contrast, we do not reveal the true realization probabilities to the players and only inform them about the underlying random process, generating the distribution over termination times, being kept constant throughout the experiment. In this treatment variation, subjects gradually learn more about the underlying process over time and hence the true distribution over termination times.[7] In order to assess the subjects' understanding of the random process, we elicited their beliefs about the distribution over termination times after the last decision round.[8]

To guarantee a clean between treatment comparison, all participants saw the same twenty-two time series in the same order. The first two were shown before the first allocation decision as examples and the remaining 20 realizations during the experiment. Final payments in the experiment were based on the points accumulated over all rounds.

[6] The formula to calculate the payoffs is not presented to the subjects in the experiment. Instead, while deciding on their allocation of resources, the payoff they receive at each possible outcome is displayed on screen in real-time and updated as soon as they reallocate any resources. This allows subjects to learn the payoffs by trial. Since we did not mention equilibrium properties in the instructions of the Blotto games, for consistency, we did not explicitly tell subjects that it is in their best interest to allocate their resources in accordance to their beliefs about the true realization probabilities, a statement that may be comparatively advantageous for the mechanism with individual-based incentives.

[7] One advantage of the use of the time series is that participants collect more information in one round of decision making compared with the classical urn experiments. Figure A.2 in the online appendix shows the (unconditional) 'empirical' distribution of the termination times generated by sampling (with replacement) from the innovations based on the two example time-series (the figure in the instructions and the example time-series on-screen) relative to the reference benchmark. The figure indicates that already before making their first decision participants can form a good impression of the process.

[8] We incentivized this task and pay a fixed amount for each 'unit' that is correctly allocated in the termination distribution. That is, subjects received the amount $\sum_{j=1}^{11} \min\{\beta_j, p_j\}$, where p is the true probability distribution and β is the expressed belief about p.

3.3 Procedures

The experiments were conducted in the experimental laboratory at Maastricht University between September 2014 and April 2018, in accordance with the peer-approved procedures established by Maastricht University's Behavioral and Experimental Economics Laboratory (BEElab). Only individuals who voluntarily entered the experiment recruiting database were invited via ORSEE (Greiner 2015), and informed consent was indicated by electronic acceptance of an invitation to attend an experimental session.

Participants operated in one of six possible treatments that varied in information condition (Baseline or Predict) and incentive mechanism (Ind, Prop or Win). They received written instructions, which they could study at their own pace. At any time, they were allowed to ask clarifying questions privately. All interactions took place anonymously via computer clients that were connected to a central server. The experiments were programmed in z-Tree (Fischbacher 2007).

We had 48 subjects participating in each of the treatments with strategic incentives and 24 subjects participating in each of the treatments with individual incentives, aggregating to 240 subjects participating in our experiment. A typical session lasted approximately 1 h and the average payoff was roughly 18 Euros (including a 3 Euro show-up fee). Sample instructions, screenshots, and time series are provided in Appendices C, D and E of the online appendix.

4 Results

We collect the sequence of twenty allocation decisions for a total of 240 subjects, giving us 24 independent observations per treatment. In this section, we first present the results for the Baseline treatments using the theoretical predictions as the null hypothesis. Then, we present the results for the Predict treatments using the theoretical predictions for the Baseline treatments as our working hypothesis. Deeper investigations into the forces that can(not) explain the main findings are relegated to the discussion section. Throughout this section, we report either two-sided Wilcoxon signed-rank or rank-sum tests for tests within and across groups respectively to investigate differences within and between treatments.

4.1 Performance in elicitation

If subjects behave in accordance to theoretical (equilibrium) prediction, in the baseline treatment with (common) knowledge of realization probabilities, they are expected to adopt allocation strategies that reveal these true realization probabilities in all three mechanisms (Ind, Prop and Win). Though, unlike in the Ind and Prop mechanism, subjects are expected to randomize over allocations in the

Win mechanism, such that true realization probabilities may not be accurately observed in realized allocations.

In order to formally assess the elicitation performance of the mechanism, we consider the accuracy of the (group) allocations ($\widetilde{\sigma}$) relative to the true distribution (p). We quantify this accuracy using the Hellinger distance (Hellinger 1909), which is defined as:

$$H(\widetilde{\sigma}, p) = \frac{1}{\sqrt{2}}\sqrt{\sum_{j=1}^{m}(\sqrt{\widetilde{\sigma}_j} - \sqrt{p_j})^2}.$$

The Hellinger distance quantifies the similarity between two probability distributions. It takes the maximal value of one in case the supports of the (group) allocation and the true distribution are disjoint and the minimal value of zero when the two distributions are identical.[9]

To evaluate the elicitation performance of the two Blotto variations, for each pair, we first average over the two individual allocations before computing the respective Hellinger distance for this pair. In the Ind treatment, in contrast, individuals do not make decisions in pairs and in order to not unjustly disfavor the Ind treatment in the comparison to the Blotto games, we correct for the performance measure by also averaging individual allocations in the Ind mechanism over randomly generated pairs before computing the Hellinger distances. This procedure ensures that we mimic the groups present in the two other treatments.[10]

Figure 4 shows the development of the average Hellinger distance between the empirical and the true distribution over the twenty periods for the three mechanisms in the baseline treatments. The figure shows that the Prop mechanism outperforms the Ind mechanism in all periods. The Win mechanism starts at a level comparable to the Prop mechanism, but its performance does not improve as fast in the beginning and within a few rounds its performance converges to that of the Ind mechanism. The test results, using the same group-level Hellinger distances as in Fig. 4, but now time averaged (over all, first ten or last ten periods), as the unit of observation, are reported in Table 1. The results show that the observed differences are

9 The Hellinger distance has been used in several applications of density estimation (cf. Birgé 1986) and computer vision analysis (cf. Zhou et al. 2010). An important advantage of the Hellinger distance over often used alternatives (such as the Kullbeck–Leibler divergence) is that it does not require absolute continuity. Another desirable property of the Hellinger distance, that we do not exploit here, is that it satisfies the triangular inequality. One notable shortcoming of the Hellinger distance, which is shared with existing alternative measures, is that it does not take into account the linear order on the domain. In order to ensure that our main findings are robust to the chosen distance measure, we replicate the analysis from the main text using the Jensen–Shannon divergence in Appendix B of the online appendix, and show that the results from both measures align closely.

10 After this pairing of individuals in the Ind treatment, the only remaining minor difference with the Blotto mechanisms is that in the latter mechanisms individuals learn after each round the number of resources the rival player has allocated to the decisive bin in that round while there is no comparable information flow between individuals for the Ind mechanism. For the Baseline treatments this is not problematic given that participants are fully informed about the true realization probabilities. For the Predict treatments we conjecture that this difference is negligible given that the information added is minute compared to the amount that individuals learn about the true realization probabilities from the feedback they receive about the process (as explained in Footnote 7).

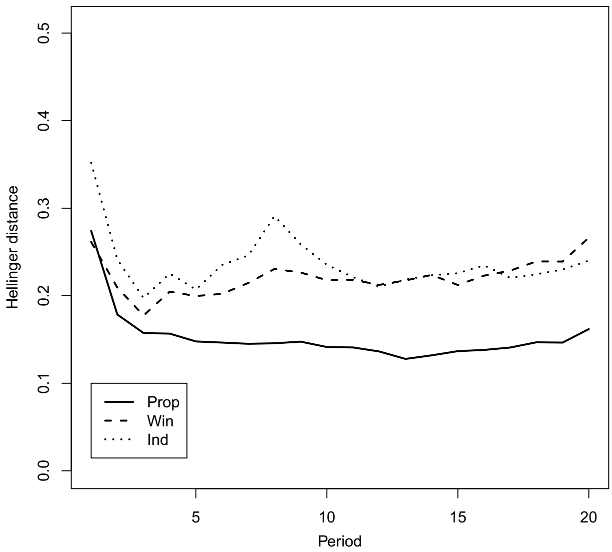


Fig. 4 Elicitation performance

Table 1 Elicitation performance

Period	Mechanisms			Comparison		
	Ind	Prop	Win	I vs. P	I vs. W	P vs. W
All	0.237	0.152	0.221	0.004	0.456	0.016
1st half	0.249	0.164	0.214	0.007	0.146	0.102
2nd half	0.225	0.141	0.228	0.004	0.830	0.002
1st vs. 2nd	0.027	0.001	0.084			

statistically significant, with Prop significantly outperforming the other two mechanisms being the most notable observation.

Another notable finding is that the performance of Win is lower than that of Prop and not statistically distinguishable from Ind. This may be a consequence of players randomizing their allocations in accordance with the equilibrium prediction in the Win treatment. If we do not account for this possibility, we may be underestimating the true elicitation ability of the Win mechanism. As a first step, before adjusting the performance measure for randomization, we assess the degree to which this is happening. We quantify the degree to which there is randomization by computing the individual Hellinger distance between a subject's average allocation and the same subject's allocation in a given period, averaged over periods.[11] Table 2 shows the

[11] To be precise, we compute for each individual the value $\frac{1}{|T|}\sum_{t\in T} H(\sigma_t^i, \tilde{\sigma}_T^i)$ with $\tilde{\sigma}_T^i = \frac{1}{|T|}\sum_{t\in T}\sigma_t^i$. For the tests we next average over players within pairs to ensure independence of observations. For the Ind mechanism we use randomly created pairs identical to those used in Table 1.

Table 2 Randomization

Period	Mechanisms			Comparison		
	Ind	Prop	Win	I vs. P	I vs. W	P vs. W
All	0.174	0.116	0.192	0.049	0.908	0.001
1st half	0.187	0.133	0.188	0.090	0.908	0.026
2nd half	0.162	0.099	0.196	0.026	0.518	0.000
1st vs. 2nd	0.042	0.000	0.317			

Table 3 Elicitation performance when controlling for randomization

Period	Mechanisms			Comparison		
	Ind	Prop	Win	I vs. P	I vs. W	P vs. W
All	0.191	0.129	0.167	0.032	0.398	0.085
1st half	0.201	0.142	0.174	0.045	0.416	0.172
2nd half	0.201	0.131	0.186	0.013	0.631	0.019
1st vs. 2nd	0.970	0.101	0.303			

resulting treatment averages and test results of the across treatment comparisons. Interestingly, players appear to vary their allocations to a similar degree in both Ind and Win, while individuals show significant less variation in the Prop treatment. This indicates that players indeed randomize more in the Win mechanism than in the Prop mechanism. To address this we correct for randomization for all three mechanisms by time-averaging the group allocations before computing the Hellinger distance. The corrected elicitation performance results are presented in Table 3.

In comparison to the other two mechanisms, the performance of the Win treatment seems to benefit more from controlling for differences in randomizing behavior, but the observed differences across treatments remain intact. From this we can conclude that in terms of elicitation ability the Win mechanism is not different from the Ind mechanism. Further, we can conclude that the higher performance of the Prop mechanism relative to the Win and Ind mechanisms can be truly attributed to the difference in incentives provided by the mechanisms.

4.2 Performance in prediction

Unlike for the baseline treatments and due to the lack of (common) knowledge of realization probabilities, we do not have a theoretical prediction concerning the ranking of the three mechanisms in the predict treatments. However, based on the example figure that the participants saw in the instructions and the example animation that was presented on screen before making their first allocation decision, they received quite detailed information about the process and the realization probabilities (recall Footnote 7). Moreover, given that all the subjects saw the same time-series in an identical order over the course of the experiment, we do not expect subjects in different treatments to form different (subjective) beliefs about the termination times. In fact, the Hellinger distance between elicited beliefs in the last period

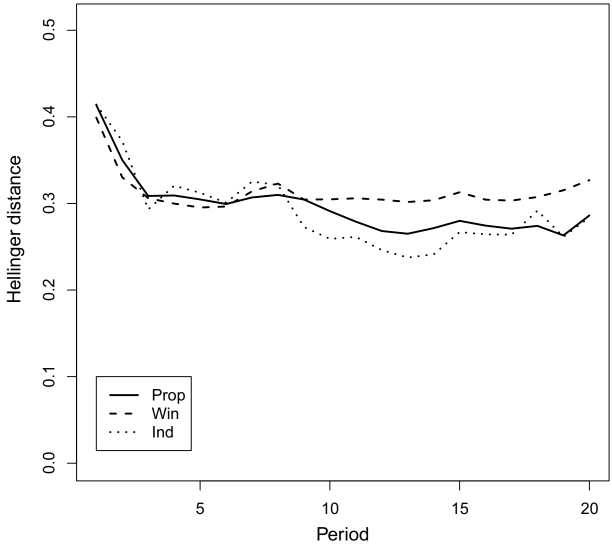


Fig. 5 Prediction performance

Table 4 Prediction performance

Period	Mechanisms			Comparison		
	Ind	Prop	Win	I vs. P	I vs. W	P vs. W
All	0.291	0.297	0.313	0.882	0.251	0.247
1st half	0.319	0.320	0.317	0.830	0.585	0.480
2nd half	0.262	0.273	0.309	0.398	0.016	0.032
1st vs. 2nd	0.000	0.000	0.345			

and the true realization probabilities are not significantly different in any pairwise comparison across mechanisms (see first row of Table 7 in the discussion). This suggests that beliefs in the Predict treatment are invariant to the mechanism.

Figure 5 shows the evolution of the average Hellinger distance between the empirical and the true distribution over the twenty periods for the three mechanisms in the predict treatments. The test results are reported in Table 4, again using group-level Hellinger distances, averaged across the relevant time periods, as the unit of observation.[12] Both the figure and the table reveal that the three mechanisms show identical levels of performance during the first half. In the second half of the experiment, Win does not display any improvement but both Ind and Prop do display a decrease in the average Hellinger distance. Consequently, both Ind and Prop

[12] In other words, Hellinger distances are computed identically to those reported in Table 1.

Table 5 Prediction performance when controlling for randomization

Period	Mechanisms			Comparison		
	Ind	Prop	Win	I vs. P	I vs. W	P vs. W
All	0.253	0.278	0.280	0.049	0.112	0.992
1st half	0.276	0.299	0.291	0.188	0.476	0.443
2nd half	0.247	0.264	0.289	0.112	0.038	0.172
1st vs. 2nd	0.042	0.000	0.812			

outperform Win in the latter half of the experiment, while they are not statistically distinguishable from each other.

Table 5 replicates the previous table but now correcting for possible randomization behavior.[13] The results from this procedure reveal that the differences across treatments are no longer that large and on average Ind now produces the lowest Hellinger distances. Statistically, Ind is better than Prop when averaging over all 20 periods and it also shows a significant edge over Win in the latter half of the experiment. Yet, taken together the results are not persuasively in favor of any single mechanism. We further elaborate upon this finding and provide additional analyses in the following section.

5 Discussion

There are several findings from the previous section that we would like to digress on. The first concerns the finding that Prop outperforms the Ind mechanism in the Baseline condition. One notable characteristic of the quadratic scoring rule (shared by many other well-known scoring rules, including the logarithmic and spherical scoring rules) is that expected payoffs are flat at the optimum, which may provide limited incentives for individuals to respond with high accuracy. In fact, moving one unit of mass from the optimal allocation decreases the expected payoff by 0.0173 points in the Ind mechanism. In contrast, moving one unit of mass from the equilibrium allocation in the Prop mechanism decreases the expected payoff by 0.5075 points at maximum. Although the latter number is a multifold of the former, also the latter value is meager compared to the maximum expected payoff of 100, indicating that also the Prop mechanism displays a flat incentive structure at the equilibrium point.

An important, and related point, is the robustness of the mechanisms presented in this paper with respect to individual risk preferences. It is known that scoring rules are only valid under the assumption of risk-neutrality and corrections for this have been proposed (Offerman et al. 2009). In the context of our experiment, risk-averse participants in the Baseline condition should submit 'flatter' distributions than is to be expected based on actual beliefs in an attempt to ensure

[13] The computation of the Hellinger distances are identical to those presented in Table 3.

Table 6 Similarity in players' allocations in Baseline (without and with correction for randomization)

	Without correction			With correction		
	Mechanisms		Comparison	Mechanisms		Comparison
Period	Prop	Win	P vs. W	Prop	Win	P vs. W
All	0.275	0.338	0.120	0.189	0.198	0.736
1st half	0.296	0.336	0.455	0.220	0.228	0.798
2nd half	0.255	0.341	0.030	0.217	0.253	0.247
1st vs. 2nd	0.042	0.877		0.623	0.152	

equal realized payoffs regardless of the outcome. The cost of doing so, in terms of 'expected payoff', is quite low in both Ind and Prop as discussed in the preceding paragraph. Similarly, a risk-averse subject who knows the realization probabilities should also not change allocations from round-to-round as this induces payoff variance. Surprisingly, the evidence on randomization behavior in the Baseline condition (Table 2) does not support this hypothesis as subjects vary their allocations significantly more in Ind compared to Prop throughout the entire experiment. Taken together with the fact that the presence of strategic uncertainty in the Blotto treatments makes it more difficult to reduce the variance of payoffs across outcomes in Prop than in Ind, it is unlikely that risk preferences are driving our findings with respect to the performance of Ind.

The second finding concerns the Win mechanism in the Baseline condition. As shown in the previous section, the inferior performance relative to Prop cannot be explained by players resorting to randomized strategies only. Another possible explanation is that the Win mechanism provides players with incentives to anti-coordinate by 'specializing' on different events, in the sense that resources placed on events on which the rival player puts more resources can be considered a waste and can better be used to increase competitiveness on other events. If players anti-coordinate, then we should see a low degree of similarity in the allocations within a group. We quantify the dissimilarity by the Hellinger distance between the allocations of the two rival players: $H(\sigma^1, \sigma^2)$. The results of this procedure are presented in Table 6. The table consists of two complementary panels, one which does not correct for randomization (left panel) and one that does (right panel). The test results in the left panel show that the dissimilarity in allocations decreases in Prop, but not in Win (comparing the 1st–2nd half). Further, the similarity in players' allocations within pairs is significantly lower in Win than in Prop during the second half. However, when correcting for randomization (right panel), these observed differences lose their significance. This means that the observed dissimilarities rather relate to differences in the realizations due to randomization, than to systematic dissimilarities in the players' underlying strategies. In case of effective anti-coordination, for instance with players randomizing over different outcomes, we should also see a significant treatment difference after correcting for randomization. Consequently, we conjecture that a significant portion of the underperformance of Win in the Baseline treatment can be attributed to ineffective randomization rather than anti-coordination.

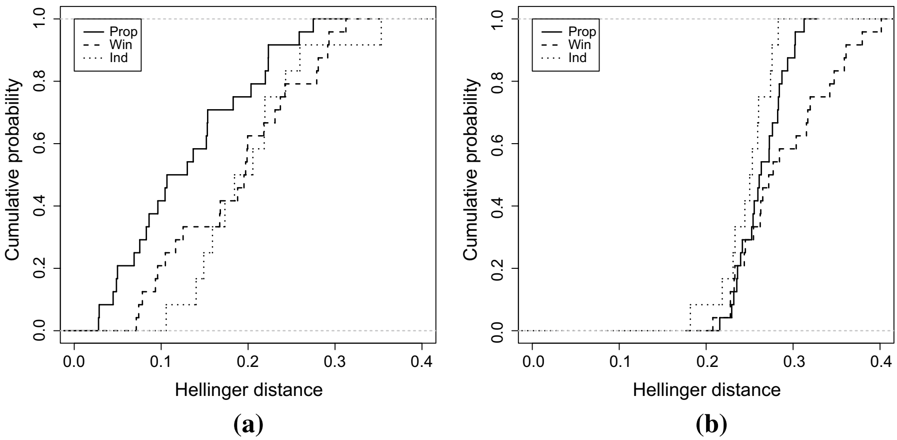


Fig. 6 Empirical CDF. **a** Baseline. **b** Predict

The third finding concerns the converging performance of all mechanisms in Predict (Table 5) relative to the Baseline (Table 3) treatment when controlling for randomization. First, the differences between Prop and Ind that were evident in the Baseline condition become much smaller in Predict and only one difference is significant. Ind outperforms Prop when comparing performance over all periods, but surprisingly not in any of the two sub-samples. Second, comparing Ind and Win, when correcting for randomization in Table 5, we find a significant difference in the second half of the experiment. Yet, taken together, the relative magnitude of the differences across mechanisms is smaller than in the Baseline condition. This suggests that once we introduce uncertainty about the environment, in this case about the distribution over termination times, subjects focus more on learning about environmental conditions and less on strategic considerations. This is a potentially important aspect to consider when implementing a competitive mechanism in practice, as it highlights the fact that, even though strategic concerns may play a role in the revelation of expectations, environmental uncertainty can overshadow the strategic component.

Comparing the empirical cumulative distribution functions of the Hellinger distance computed over groups, and correcting for randomization effects, depicted in Fig. 6, clear differences in performance are visible in the Baseline treatment while these are much smaller in the Predict treatment. In fact, when comparing the distributions of Hellinger distances of the Ind and Prop treatments in the Predict condition using a Kolmogorov-Smirnov test, we fail to reject equality of these distributions ($p = 0.485$). We also compare the variance of the two treatments with a Levene test and also there we fail to reject equal variances ($p = 0.891$). This suggests that there is no substantial performance differential between the Ind and Prop treatments in Predict and hence, the game mechanism using proportional payoffs does not systematically perform worse than the scoring rule based mechanism.

Fourth, a possible concern with our strategic mechanisms is that allocations are not free from beliefs about the allocations of the opponent. While a game-based method cannot completely eliminate this strategic uncertainty, we are able to

Table 7 Hellinger distance between the beliefs elicited in the last period and the true distribution and the actual allocations

Comparison	Mechanisms			Comparison		
	Ind	Prop	Win	I vs. P	I vs. W	P vs. W
$H(\beta, p)$	0.263	0.268	0.306	0.971	0.403	0.367
$H(\beta, \sigma_{11-20})$	0.159	0.241	0.332	0.002	0.000	0.000

compare the beliefs elicited in the last period in the Predict treatment to the actual allocations under the three payoff rules. The results of this comparison are presented in Table 7. The second row of the table shows the average Hellinger distance between the beliefs directly elicited and the average individual allocations in the last ten periods (such that the unit of observation is on individual level; giving 48 observations for Prop and Win and 24 for Ind), and the comparison between mechanisms (with p-values from the Wilcoxon rank-sum tests reported). As expected, we find that allocations are closest to beliefs in the Ind treatment and furthest away in the Win treatment. This suggests that in Ind, being free from strategic concerns, individuals are most truthfully revealing their beliefs. However, these beliefs do not translate into better forecasting performance, as shown in Tables 4 and 5. Moreover, as noted before, beliefs in Ind are not significantly closer to the true realization probabilities when compared to either Prop or Win (first row of Table 7). Taken together, it implies that while beliefs and strategies are more similar in Ind, they are not better than those in any of the other treatments.

Finally, it is important to put the quadratic scoring rule employed in our Ind treatment in perspective. While the mechanism is shown to work well in Harrison et al. (2017), it does not appear to be performing as well in our experiment as it did in theirs. There is an important difference in the experimental design and setup that may cause some of this difference. In the relevant treatments of Harrison et al. (2017) subjective beliefs about an already realized event are elicited, and the subjects' uncertainty concerns the identification of this realized event.[14] In our experiment there is considerably more uncertainty in the sense that subjects are asked to estimate the realization from an unknown process about which they only possess limited information. We believe that our findings in the Predict treatment are more relevant for applications focused on predicting future, uncertain, events rather than collecting information on the situation at present, where proper scoring rules have been shown to perform well.

[14] Specifically, subjects are asked to estimate the proportion of red balls in an urn which is shown to the participants.

6 Conclusion

We present two variations of the Colonel Blotto game with two players placing resources on possible future events: one variation uses a proportional-prize rule, the other a winner-takes-all rule. When realization probabilities are common knowledge to the players, theory predicts both players to reveal these probabilities via their allocation of resources under the proportional-prize rule, while they randomize their allocations in a way to fully reflect the probabilities in expectation under the winner-takes-all rule. Hence, both variations possess elicitation capacity, and we test this capacity via a laboratory experiment using a proper scoring rule as benchmark. We find that the Blotto game augmented with the proportional-prize payment rule outperforms both other mechanisms, providing evidence for this game being a good mechanism to elicit objective beliefs.

Since mechanisms that are developed to elicit individual expectations are also applied in forecasting environments, we repeated the experiments in a setting where players do not know the true realization probabilities and gradually gather more information during the course of the experiment. Here we find the Blotto game with proportional-prize payoffs to not be statistically distinguishable from the mechanism using individual incentives via a quadratic scoring rule. This suggests that, while there is strategic uncertainty present in the game, as opposed to the scoring rule, this does not translate into worse performance. Taken together, this reinforces the evidence that games can be used to elicit expectations successfully and offers a viable alternative to a scoring-rule based approach.

References

Ames, C. (1984). Achievement attributions and self-instructions under competitive and individualistic goal structures. *Journal of Educational Psychology*, *76*(3), 478–487.

Avrahami, J., & Kareev, Y. (2009). Do the weak stand a chance? Distribution of resources in a competitive environment. *Cognitive Science*, *33*, 940–950.

Avrahami, J., Kareev, Y., Todd, P. M., & Silverman, B. (2014). Allocation of resources in asymmetric competitions: How do the weak maintain a chance of winning? *Journal of Economic Psychology*, *42*, 161–174.

Baillon, A. (2017). Bayesian markets to elicit private information. *Proceedings of the National Academy of Sciences*, *114*(30), 7958–7962.

Birgé, L. (1986). On estimating a density using Hellinger distance and some other strange facts. *Probability Theory and Related Fields*, *71*(2), 271–291.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*, 1–3.

Carpenter, J., Holmes, J., & Matthews, P. H. (2008). Charity auctions: A field experiment. *Economic Journal*, *118*(525), 92–113.

Carvalho, A. (2016). An overview of applications of proper scoring rules. *Decision Analysis*, *13*(4), 223–242.

Chen, K.-Y., & Plott, C. R. (2002). *Information aggregation mechanisms: Concept, design and implementation for a sales forecasting problem*. New York City: Mimeo, Inc.

Chowdhury, S., Kovenock, D., & Sheremeta, R. (2013). An experimental investigation of Colonel Blotto games. *Economic Theory*, *52*(3), 833–886.

Cowgill, B., Wolfers, J., & Zitzewitz, E. (2009). *Using prediction markets to track information flows: Evidence from Google*. New York City: Mimeo, Inc.
Duffy, J., & Matros, A. (2017). Stochastic asymmetric Blotto games: An experimental study. *Journal of Economic Behavior & Organization*, *139*, 88–105.
Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., et al. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences*, *112*(50), 15343–15347.
Figlewski, S. (1979). Subjective information and market efficiency in a betting market. *Journal of Political Economy*, *87*(1), 75–88.
Fischbacher, U. (2007). zTree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, *10*(2), 171–178.
Forsythe, R., Nelson, F., Neumann, G. R., & Wright, R. (1992). Anatomy of an experimental political stock market. *American Economic Review*, *82*(5), 1142–1161.
Friedman, L. (1958). Game-theory models in the allocation of advertising expenditures. *Operations Research*, *6*(5), 699–709.
Gillen, B. J., Plott, C. R., & Shum, M. (2017). A pari-mutuel-like mechanism for information aggregation: A field test inside Intel. *Journal of Political Economy*, *125*(4), 1075–1099.
Greiner, B. (2015). Subject pool recruitment procedures: Organizing experiments with ORSEE. *Journal of the Economic Science Association*, *1*(1), 114–125.
Gross, O. (1950). *The symmetric Blotto game. RAND RM-2242*. Santa Monica: RAND Corporation.
Harrison, G. W., Martinez-Correa, J., Swarthout, J. T., & Ulm, E. R. (2017). Scoring rules for subjective probability distributions. *Journal of Economic Behavior & Organization*, *134*, 430–448.
Hellinger, E. (1909). Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen. *Journal für die reine und angewandte Mathematik*, *136*, 210–271.
Kovenock, D., & Roberson, B. (2015). *Generalizations of the General Lotto and Colonel Blotto games*. Working paper.
Laslier, J. F. (2002). How two-party competition treats minorities. *Review of Economic Design*, *7*(3), 297–307.
Niederle, M., & Vesterlund, L. (2007). Do women shy away from competition? Do men compete too much? *Quarterly Journal of Economics*, *122*(3), 1067–1101.
Offerman, T., Sonnemans, J., van de Kuilen, G., & Wakker, P. P. (2009). A truth serum for non-Bayesians: Correcting proper scoring rules for risk attitudes. *Review of Economic Studies*, *76*(4), 1461–1489.
Peeters, R., & Wolk, L. (2017). Eliciting interval beliefs: An experimental study. *PLoS ONE*, *12*(4), e0175163.
Plott, C. R., Wit, J., & Yang, W. C. (2003). Parimutuel betting markets as information aggregation devices: Experimental results. *Economic Theory*, *22*(3), 311–251.
Prelec, D. (2004). A Bayesian truth serum for subjective data. *Science*, *306*(5695), 462–466.
Robson, A. (2005). *Multi-item contests*. Working paper.
Thaler, R., & Ziemba, W. T. (1988). Parimutuel betting markets: Racetracks and lotteries. *Journal of Economic Perspectives*, *2*(2), 161–174.
Thomas, C. D. (2018). *N*-dimensional Blotto games with heterogeneous battlefield values. *Economic Theory*, *65*(3), 509–544.
Zhou, X., Yu, K., Zhang, T., & Huang, T. S. (2010). Image classification using super-vector coding of local image descriptors. ECCV 2010. *Lecture Notes in Computer Science*, *6315*, 141–154.

Affiliations

Ronald Peeters[1] · **Leonard Wolk**[2]

Ronald Peeters
ronald.peeters@otago.ac.nz

Leonard Wolk
l.wolk@vu.nl

1 Department of Economics, University of Otago, Dunedin, New Zealand

2 Department of Finance, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands