SELF-RESOLVING INFORMATION MARKETS: AN EXPERIMENTAL CASE STUDY

Kristoffer Ahlstrom-Vij
Department of Philosophy
Birkbeck College
30 Russell Square
London WC1B 5DT
UNITED KINGDOM
k.ahlstrom-vij@bbk.ac.uk

Nick Williams
Dysrupt Labs
Scottish House
Level 4, 90 William St
Melbourne, Victoria, 3000
AUSTRALIA
nickwilliams@dysruptlabs.com

Abstract: On traditional information markets (TIMs), rewards are tied to the occurrence (or non-occurrence) of events external to the market, such as some particular candidate winning an election. For that reason, they can only be used when it is possible to wait for some external event to resolve the market. In cases involving long time-horizons or counterfactual events, this is not an option. Hence, the need for a *self-resolving* information market (SRIM), resolved with reference to factors internal to the market itself. In the present paper, we first offer some theoretical reasons for thinking that, since the only thing that can be expected to be salient to all participants on a SRIM is the content of the question bet on, a convention will arise of taking that question at face value, and betting accordingly, in which case trading behaviour on SRIMs can be expected to be identical to that on TIMs. This is the *'face value' hypothesis*. If this hypothesis holds, SRIMs have the potential of incorporating the accuracy of TIMs while shedding their limitations in relation to long-term predictions and the evaluation of counterfactuals. We then report on a laboratory experiment that demonstrates that trading behaviour *can* indeed come out highly similar across SRIMs and TIMs. As such, the study can be thought of as an experimental case study on SRIMs. Finally, we discuss some limitations of the study, and also points towards fruitful areas of future research in light of our results.

Key words: Information markets; Prediction markets; Self-resolving information markets; long-term information markets; Counterfactuals; Experimental information markets

1. Introduction

Information markets, also known as prediction markets, are markets for placing bets on future or otherwise unknown events.[1] On what we might call *traditional* information markets (TIMs), rewards are tied to the occurrence (or non-occurrence) of events external to the market, such as some particular candidate winning an election, a central bank raising or lowering the interest rate by some specific amount, and so forth. This creates clear incentives to bet in accordance with what one takes the relevant facts to be. If one does, and one is right, one sees a good return. As a result, those in the

[1] We prefer the term 'information markets' to the more popular 'prediction markets,' since the latter gives the mistaken impression that such markets are restricted to the domain of forecasting, despite being at least in principle capable of assigning probabilities to a range of unknown events, not restricted to future ones, including events in the past or (as we shall see) counterfactual events.

know have good reason to reveal their knowledge on TIMs, and will in so doing profit—by way of a financial return (on real-money markets) or simply through the gratification of being shown to be right (on play-money markets)—from the liquidity provided by those who happen to be less informed on the relevant matters (Ahlstrom-Vij 2016).

Consequently, it should come as no surprise that the price signals arising on such markets, if interpreted as probability assignments, generally constitute good approximations of the likelihood of events in a wide range of areas (Hahn and Tetlock 2006), including politics (Berg and Rietz 2014; Berg, Nelson, and Rietz 2008; Forsythe et al. 1998), sports (Luckner, Schröder, and Slamka 2008; Deschamps and Gergaud 2007; Debnath et al. 2003), business (O'Leary 2011; Chen and Plott 2002; Spann and Skiera 2003), medicine and health care (Rajakovich and Vladimirov 2009; Polgreen et al. 2007; Mattingly and Ponsonby 2004), and entertainment (McKenzie 2013; Pennock et al. 2001).[2]

At the same time, for reasons that will be discussed in Section 2, there is an obvious limitation to TIMs: they can only be used when it is possible to wait for some external event to resolve the market. In some cases, such as those involving very long time-horizons or counterfactual events, this is not an option. This motivates the introduction, in Section 3, of a type of market that has not received sufficient attention in the literature: a *self-resolving* information market (SRIM), resolved with reference to factors internal to the market itself. While 'the fundamentals of the concept [of a TIM] have been sufficiently understood' (104), as noted by Horn and colleagues (2014) in their extensive literature review, SRIMs have received almost no attention in the literature.[3] We therefore start out below by offering some theoretical reasons for thinking that SRIMs will take the form of a particular type of (impure) coordination game. Specifically, since the only thing that can be expected to be salient to all participants on a SRIM is the content of the question bet on, a convention will arise of taking that question at face value, and betting accordingly, in which case trading behaviour on SRIMs can be expected to be identical to that on TIMs. We refer to this as the *'face value' hypothesis*.

If this hypothesis holds, it would be highly significant, as it would mean that SRIMs have the potential of incorporating the accuracy of TIMs while shedding their limitations in relation to long-term predictions and the evaluation of counterfactuals. This is the motivation for the laboratory experiment discussed in Sections 4 and 5—an experiment we ran in order to add to a virtually non-existent, experimental literature on SRIMs.[4] While underpowered for purposes of statistical inference, our results demonstrate that trading behaviour *can* come out sufficiently similar across SRIMs and TIMs. As such, the study can be thought of as an experimental *case study* on SRIMs. Finally, Section 6 discusses some limitations of the study, and also points towards fruitful areas of future research on SRIMs in light of the study's results.

2. Challenges for Traditional Information Markets

Many things that we want to predict take place far into the future. For example, consider the variety of long-term factors on which questions about the severity of climate change turns, such as the temperature dependent movements of seawater, in turn affecting water density, or the release of

[2] For comprehensive accounts of relevant literature, see Tziralis and Tatsiopoulos (2007), covering the period of 1990 through to 2006, and Horn et al. (2014), covering the period of 2007 until 2013, as well as Klingert (2017) for an analysis of the publications with the highest impact on the information market literature.

[3] As we shall see, the one exception here is Abramowicz (2007).

[4] To the best of our knowledge, there has only been one experimental study on such markets, by Espinoza et al. (ms.) in the context of risk and vulnerability studies, the results of which are confidential on national security grounds.

methane from arctic tundra. Being able to predict changes in factors such as these is crucial to managing the effects of climate change. Given the impressive track record of information markets, we would if at all possible want a way to utilise their predictive power in these and similar contexts.

However, TIMs run into problems when implemented for purposes of predicting events far into the future. Specifically, Antweiler (2012) suggests that there are two main challenges for long-term information markets, as in markets with a time-horizon measured not in weeks or months but in years. The first challenge is that the long time-horizon will result in low liquidity, owing to the attention of traders fading over time (assuming they get involved to begin with). The second problem is that the opportunity costs presented by traders tying up their money for a long time, and thereby not being able to earn a return elsewhere, will make such markets unattractive.

Antweiler argues that the latter problem about opportunity costs can be overcome by the market maker compensating traders through separate investments. But there is of course another way to avoid that problem: by having the market operate on the basis of play money. Several results are relevant here. Servan-Schreiber and colleagues (2004) found no difference in accuracy between real-money and play-money markets, while Rosenbloom and Notz (2006) found slightly higher levels of accuracy for real-money markets. McHugh and Jackson (2012) were able to explain these seemingly inconsistent results by showing that context matters. In particular, in markets dedicated to politics and sports, there is no accuracy difference between real- and play-money markets.

That said, opting for a play-money market doesn't address the first challenge identified by Antweiler: that the market will suffer from low liquidity, owing to there being too great a discrepancy between the long time-horizon of the market and the short attention span of potential traders. One way of meeting this challenge, however, is offered by Graefe and Weinhardt (2008), who resolve contracts on long-term markets with reference to the outcome of separate markets consisting exclusively of expert traders. This manner of resolution drastically reduces the market's time-horizon, and thereby avoids not only the problems of opportunity costs, but also that of waning interests on the part of potential traders over time.

However, the limitations of Graefe and Weinhardt's approach are fairly obvious. For one thing, it requires us to always run two markets, with two sets of traders, instead of one. For another—and more importantly—their approach is only feasible in a context where we already have a good sense of who the experts on the relevant matters are. And in cases where we do, it is less clear why we would be looking to set up an information market to begin with. To see why, note that, when we talk about 'experts' here, we have in mind the people who happen to be informed on the topic at hand. This might not coincide with the people who have, for one reason or the other (including existing power structures that might not track genuine competency), been *designated* experts. This is why information markets are helpful, as one of their main attractions is that they enable us to harness the insights of experts in contexts where we don't necessarily know who the experts are, but where we trust that, whoever they are, they will reveal themselves on the market. As such, information markets solve what is in many contexts a notoriously difficult *expert identification problem*.

Consequently, even in light of Antweiler's and Graefe and Weinhardt's suggestions, the challenge associated with setting up a market that successfully predicts events far into the future very much remains.

TIMs also run into problems when trying to evaluate counterfactuals. Indeed, here we face an even more formidable challenge: If the main problem when it comes to predicting events far into the future using TIMs is that traders are unwilling to wait around until the point in time where the market resolves, the problem for markets trading in counterfactual events is that no such point in time exists. Consider an example: *Had Russia not interfered in the 2016 US Presidential election, Hillary Clinton would have won*. This statement has a truth-value, and we might be interested on getting a sense of

what it is, in so far as we want to understand whether Russian interference changed the outcome of the election. But since the antecedent of the counterfactual will never come out true—if Russia interfered in the 2016 election (and that much seems established), it will never be the case that they did not—it is simply not an option to set up a market that is settled by waiting for the antecedent to come true, and then evaluating the consequent.[5] For that reason, it is arguably impossible to set up a TIM concerned with counterfactual events such as these.

Here, we need to be careful not to confuse counterfactuals with (indicative) conditionals, such as *If Hillary Clinton runs again, she will lose*. Setting aside the problems considered a moment ago in relation to long time-horizon, we can here set up a TIM using *conditional* contracts. As discussed by Hanson (2013), such contracts are traded conditional on the antecedent of the relevant conditional statement (e.g., Clinton runs again). If the event does not come to pass, the trades are called off, and everyone is paid back whatever they have staked. Nevertheless, as Hanson points out, until such a time, trading on the relevant markets 'gives us speculator estimates on the consequences of events that never actually happen; until speculators know an event won't happen, they can have incentives to accurately estimate its consequences' (159). It follows from this that there cannot be any incentives of such a kind in relation to counterfactual events, as the antecedents by definition will never come out true over time, and traders know this.

Still, the fact that important decisions hang on our being able to make accurate judgments about the probability of events far into the future, and about the likely truth-value of counterfactual statements, makes it worthwhile to ask whether there is still a way to harness the power of information markets in such contexts. In particular, given the challenges we have seen arise for TIMs, the question arises whether there is a way to tie rewards on information markets, not to the occurrence of some external event, but instead to events internal to the market. This would solve the problem of opportunity costs and do away with the need for a time-horizon outstripping people's attention spans on long-term markets, and also avoid the need for external resolution on a market dealing with counterfactual events.

3. Self-Resolving Markets and the 'Face Value' Hypothesis

One way to set up such a *self-resolving* information market is by having markets be settled on the basis of the final market price at some pre-specified time, unknown to the participants. So, instead of rewarding participants to the extent that their bets have helped push the price signal towards the 'true' value—which on a binary market will be either 0 or 1—a self-resolving market will reward participants to the extent that they have pushed the price signal towards whatever price the market closes at. The challenge for anyone wishing to implement such a market, however, is that we currently lack any reason to construe a person's willingness to place any particular bet on such a SRIM as revealing an estimation of the probability of any event external to the market.

To appreciate the force of this challenge, it is helpful to consider the fact that, on a TIM, there is a very obvious way in which trading behaviour will be disciplined by external events. No matter what the market price is at present, the informed trader can rest assured that he or she will eventually be proven right, and compensated accordingly. Indeed, the farther off the market price is at the point that

[5] This is not to deny that future (or indeed past) events relating to other elections and attempts at influencing elections might offer *evidence* regarding whether the counterfactual statement in question is likely true or false, by making more or less likely claims about the underlying causal mechanisms. What is being denied is that such events can in any straightforward sense *settle* the relevant markets, in the manner that TIMs are traditionally settled by the external events mentioned in the contracts traded.

the informed trader enters the market, the more handsome her eventual reward. By contrast, once market rewards are no longer tied to external events, informed traders can no longer take comfort in the fact that they will eventually be proven right about the external event (supposedly) bet on—because for all they know they might not. Consequently, if enough (potentially misinformed) traders take the market in some particular direction, the informed trader might have no choice but to bet, not against the background of her best estimate of the likelihood of the external event, but in accordance with her expectations about where the market will eventually end up at the point of self-resolution. The worry, then, is that there are reasons to believe that SRIMs will simply take the form of a 'Keynesian beauty contest,' where we, as Keynes put it, 'devote our intelligences to anticipating what average opinion expects the average opinion to be' (Keynes 2015/1936: 211), and in so doing end up making judgments that might very well be completely divorced from any considerations external to those opinions.

That said, there are also some considerations suggesting that this is *not* what will happen. To begin with, some highly successful self-resolved markets already exist, namely stock markets. Pay-offs on stock markets are *not* determined through some great closing event, where the 'correct' value of each stock is revealed, but are a function of continuous bets on what people will be prepared to pay for what in the future. This is a form of self-resolution: pay-offs are a function of a feature internal to the market, namely market price. Despite Keynes's concerns—after all, Keynes's beauty contest was supposed to illustrate a worry he had about stock markets—trades are by convention often grounded in fundamentals. We treat good fundamentals as having a positive impact on share prices, and expect that others will do the same—and shares rise as a result. So, while internally resolved, stock markets offer clear incentives to those in the know to reveal their knowledge by trading.

We want to stress that we do *not* believe that SRIMs will function exactly like stock markets. Still, it's helpful to keep in mind the convention to factor in fundamentals on stock markets when considering our hypothesis. Because what we hypothesize is that people on SRIMs will bet with an eye towards the relevant external facts on SRIMs on account of such markets developing into a particular type of *coordination game* (Abramowicz 2007; see also Schelling 1960 and Lewis 1969). How so? For one thing, since the market price is a function of the sum of bets, this creates clear incentives to bet in accordance with expectations of how other people will be trading. However, note that SRIMs aren't *pure* coordination games; partly, they're also games of *conflict* (Ahlstrom-Vij ms.). Specifically, on the type of market scoring rule used on many information markets, high rewards are given to those who take high risks by moving the market a significant distance towards the 'correct' value (e.g., Hanson 2007). In the case of SRIMs, that means being the first person to predict what people will be coordinating on, and thereby getting a first mover advantage.

Of course, successfully predicting the bets of others requires making certain assumptions about what considerations they are bringing to bear on their bets. So, what should participants assume on that score? Appreciating what type of game they are playing, they will realize that successful coordination requires the considerations to be *salient* to everyone involved. This brings us to our hypothesis:

> *The 'Face Value' Hypothesis (FVH)*: Since the only thing that can be expected to be salient to all participants on a SRIM is the content of the question bet on, a convention will arise of taking that question at face value, and betting accordingly.

The FVH is significant because, if it is true, we can expect similar trading behaviour on SRIMs as on TIMs—in both cases, people will bet with reference to what they take the relevant facts to be. If that is so, we can moreover construe a willingness to place a bet on a SRIM as revealing an estimation

about the likelihood of the relevant external event, in much the same way that we do on a TIM. And, importantly, given that TIMs tend to generate accurate outputs, the same will thereby go for SRIMs, which will then have all the benefits of TIMs while lacking a significant drawback in not requiring external resolution.

4. Evaluating the Face Value Hypothesis

How plausible is the FVH? Aforementioned argument regarding SRIMs as (impure) coordination games gives us *some* theoretical reasons to assume that it holds, but on their own these reasons are too weak to support a well-grounded expectation about trading behaviour. Add to this the fact that the experimental literature on SRIMs is, as already noted, extremely thin. To the best of our knowledge, there has only been one experimental study on such markets, by Espinoza and colleagues (ms.) in the context of risk and vulnerability studies, the results of which are confidential due to incorporating information sensitive on national security grounds, but supposedly consistent with the FVH (Espinoza, personal communication).

Using a platform developed by Dysrupt Labs, we therefore ran an experimental study, the object of which was two-fold: first, to remedy the complete absence of any publicly available, experimental data on SRIMs; and, second, to evaluate the FVH specifically by comparing trading behaviour on markets identical in all respects, save for the fact that half were externally resolved and half were self-resolved. If the FVH holds, we should expect such behaviour to be very similar across relevant metrics (more on these below).

4.1. Methodology

Six participants—four students at the University of Kent in Canterbury, UK (where one of the experimenters was based at the time), and two professionals in Melbourne, Australia (where the other experimenter was based)—were recruited on the basis of a participant information sheet introducing the study as follows:[6]

> You've been invited to participate in abovementioned study on information markets. An information market is a market for placing bets on the occurrence or non-occurrence of future or otherwise unknown events. The price signals arising on such markets tend to offer good approximations of the likelihood of the events bet on. So, for example, if the market price for a contract worth £1 if the Conservatives win an outright majority in the upcoming general election in the U.K. is 80p, the market can be taken to suggest that the event in question is 80% likely. But there's an obvious limitation to such markets: they can only be used when waiting for some external event to resolve the market is an option. Sometimes it isn't, such as in the case of events far into the future, or counterfactual events. This raises the question whether there's a way to tie rewards, not to the occurrence of some external event, but instead to events internal to the market, making for a self-resolving market. That's the question we hope to answer as part of this study.

On the day of the study (July 19, 2017), the participants were randomly allocated into groups of three to either a TIM or an otherwise identical SRIM. We did this over the course of ten questions, or 'rounds,' with a new random allocation of participants for each round. While randomly allocated to either a SRIM or a TIM for each question, the participants were aware of what type of market they were participating in, once allocated, owing to each market being marked either **[TRADITIONAL MARKET]** or **[SELF-RESOLVING MARKET]** at the top of the interface. This is because we wanted to know

[6] Full participant information sheet, signed consent forms, and institutional decision on ethical approval from the University of Kent, Canterbury, are all available upon request.

how participants bet on markets when knowing full well how their bets were being rewarded. For this purpose, each participant was also asked to watch a 9-minute video tutorial (available on request) prior to taking part in the study, explaining how to navigate the platform and spending two minutes explaining the difference between TIMs and SRIMs.

Each market question concerned the probability of drawing a black ball out of an urn with a particular distribution of black and white balls, set by the experimenters to a value unknown to the participants. On the platform, the participants in traditional markets where prompted as follows:

> **[TRADITIONAL MARKET]**
> **If a ball were to be drawn from the urn, what's the probability that it would be a black ball?**
> Imagine that there is an urn with black and white balls in it. If we were to draw a ball from the urn at random, what's the probability that a black ball would be drawn?[7]

The prompt for the self-resolving markets was identical, save for those markets being designated as self-resolving as opposed to traditional in the heading.

At one-minute intervals, participants received different samples randomly drawn (with replacement) from the distribution (e.g., 'A white ball was just drawn from the urn, before being put back'). They could then choose to factor in those samples—together with whatever samples they might take the market to be aggregating through other participants' bets—in their bets, using a limited number of points (1,000) allocated at the start of each round.

The markets used a standard version of Hansons's market scoring rule (e.g., Hanson 2007), sequentially rewarding participants to the extent that they, through their bets, move the price signals towards the correct value, less any reward that is due to past participants moving the market in the same direction. The rewards structure on TIMs was explained to the participants as follows:

> How will my bets be rewarded?
> If there are far more black than white balls in the urn, the probability of drawing a black ball is high; if there are far more white than black, then it's low. When the market closes, bets will therefore be measured against the actual proportion of black balls (i.e., 0-100%) in the urn from which you've received samples—although you won't find out what that proportion is until the end, when the market has closed.

This explanation appeared right below the above prompt on each traditional market. The corresponding explanation for the SRIMs was as follows:

> How will my bets be rewarded?
> If there are far more black than white balls in the urn, the probability of drawing a black ball is high; if there are far more white than black, then it's low. But here we imagine that there's no way of knowing the proportion of black (or white) balls in the urn. All we can do is draw samples from it and send them over to you. When the market closes, bets will therefore be measured, not against the unknown proportion of black balls, but against the market price at the time of closing. This price represents the market's judgment on the probability that a black ball would be drawn, in light of the samples aggregated on the market through the bets placed by you and others.

[7] This prompt might be taken as a counterexample to our claim from earlier (in Section 2) to the effect that we cannot evaluate counterfactuals by way of TIMs. But note that what the question ultimately asks about is a probability, which in turn is interpreted as an actual (not counterfactual) proportion, as per the rewards prompt for the TIMs below.

In other words, our aim was to make the two parallel markets identical—with the same underlying distribution of black balls, same samples being communicated to the participants, and so forth—save for the manner in which the bets were rewarded.

Every market resolved at a random point in time between 60 and 120 seconds after the fifth and final sample had been sent out. A crucial aspect of the design was that participants were completely unaware of when the market would close; indeed, they were not even informed about aforementioned time interval. This was important—and will likely be important for the design of successful SRIMs more generally—since we wanted to avoid participants on the SRIMs trying to increase their reward by placing a large bet at the very final moment of the market, thereby guaranteeing that they, being the final mover of the market, would be correct about where the market ends up at closing. There was of course nothing that prevented them from still *trying* to time their final bets in that manner, although not without running the risk of being wrong about exactly when the market is to close, and as such potentially being punished by further bets by others. In that respect, there was a fairly strong incentive against such timed bets.

Each participant was allocated a £10 incentive for participating in the study. For any individual participant, that sum could be increased when augmented with the pay-outs of successful trading on the markets—i.e., trading that pushed the market price towards the value representing the underlying distribution of black balls, in the case of the TIMs, and towards whatever the market price ends up being at the time of resolution, for the SRIMs—with the total gains and losses across markets offering an exchange rate between market points and money of 100 points per £1. However, participants could also lose their incentives if they engaged in poor trading. This was to discourage destructive trading where participants, knowing they will be guaranteed £10, start betting without any regard for potential losses. (Under no circumstances could the participants walk away owing any money.)

4.2. Results

As for our findings, it is important to note at the outset what we did *not* take the study to be able to demonstrate. Given the sample size (twenty markets in total, with ten markets per condition), the power was too low for us to detect any statistically significant differences between the two conditions (i.e., SRIM and TIM). Consequently, even if there are in fact differences between trading behaviour on SRIMs and TIMs *in general*, that is not something we could necessarily expect to pick up on within the framework of the present study.

So, what can the study help demonstrate? Whether trading behaviour *can* come out sufficiently similar across the two conditions, as investigated through an experimental *case study* on a particular (small) sample of TIMs and SRIMs. Anyone sceptical about such self-resolution, perhaps on account of suspicions (outlined in Section 3 above) that SRIMs will quickly collapse into Keynesian beauty contests, should expect not, and should correspondingly be surprised if the similarity comes out high—or, failing that, at the least be sufficiently intrigued to agree that the matter warrants further investigation.

Similarity with respect to *what*? The metrics used were as follows:

4.2.1. Comparative Volatility

If participants on a SRIM are coordinating on the question asked and taking it at face value, in much the same way as they would on a TIM, the volatility on SRIMs should be roughly equivalent to that on TIMs. Following the established practice of measuring stock volatility by way of standard deviation, we used the same measure when determining information market volatility. More specifically, using the prices at every second from the first bet as our price points, volatility was calculated as follows:

$$market\ volatility = \sqrt{\frac{\sum_{i=1}^{n}(price\ point_i\ -\ mean\ price)^2}{n-1}}$$

Using this measure, the average volatility of the TIMs and SRIMs in the study was not only relatively low, but as it turns out also identical between the two conditions: .17.

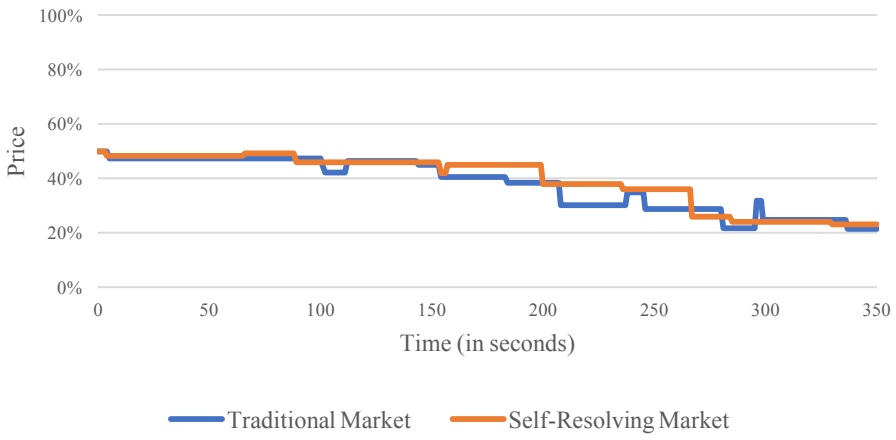
4.2.2. Market Profile Similarity

If trading behaviour on TIMs and SRIMs is for all practical purposes identical—and, in particular, if people under both conditions are trading with reference to the external facts referenced in the question asked, as per the FVH—the market profiles developing under the two conditions can be expected to be similar. To measure similarity of market profiles, we used an inversed squared-distance measure, where 'p_t^a' corresponds to the price at time *t* for market *a*, and 'p_t^b' the same for market *b*, with *i* being incremented every second of the market, starting with the point at which a bet had happened on both markets:

$$market\ profile\ similarity\ score = 1 - \sum_{i=1}^{t}\frac{(p_t^a - p_t^b)^2}{n}$$

Two markets whose price graphs match each other perfectly will have a similarity score of 1, while two markets whose prices constantly are at the opposite ends will have a similarity score of 0. As for operationalizing the idea of two market profiles being 'similar', it's helpful to consider the markets in the study's first round by way of a benchmark:

GRAPH 1. MARKET PROFILES FOR ROUND 1



Prices given for 'Yes' on the question 'If a ball were to be drawn from the urn, what's the probability that it would be a black ball?'

The two markets in Graph 1 have what would have to be considered similar market profiles, and on the above measure the similarity score does indeed come out high: .988. Comparing that score to the average score across all ten rounds—.975—makes clear that the markets generally came out similar throughout the study.

4.2.3. Final Price Similarity

While it is interesting to note the high average market profile similarity score, we were to some extent more interested in whether the respective markets would end up at the same (final) price, whether they took the same routes there. For that reason, we also calculated the similarity score specifically for the final market prices as follows, where 'p^a' corresponds to the final market price on market a, and 'p^b' the same for market b:

$$final\ market\ price\ similarity\ score = 1 - (p^a - p^b)^2$$

Here, too, it's helpful to use the markets in Graph 1 as a benchmark, as they closed at 21.54% in the case of the TIM and 22.93% in the case of the SRIM—closing prices that would have to be considered similar. Indeed, using the above measure, the final market price similarity score for round 1 came out to .999. As can be seen in Table 1, round 1 was not an outlier in this respect, given that the average final market price similarity score across the experiment's ten rounds came out to .966, with an SD of .005:

TABLE 1. FINAL MARKET PRICE SIMILARITY

ROUND	MARKET TYPE	PRICE	SIMILARITY
1	Traditional	21.54%	.999
	Self-resolving	22.93%	
2	Traditional	85.64%	.992
	Self-resolving	94.85%	
3	Traditional	27.60%	.969
	Self-resolving	9.97%	
4	Traditional	1.75%	.994
	Self-resolving	9.51%	
5	Traditional	19.43%	.970
	Self-resolving	2.12%	
6	Traditional	23.67%	.995
	Self-resolving	30.49%	
7	Traditional	88.24%	.982
	Self-resolving	74.86%	
8	Traditional	79.55%	.992
	Self-resolving	70.37%	
9	Traditional	72.32%	.770
	Self-resolving	24.32%	
10	Traditional	92.92%	.998
	Self-resolving	97.74%	
	AVERAGE SIMILARITY:		.966
	SD:		.005

All **prices** given for 'Yes' on the question 'If a ball were to be drawn from the urn, what's the probability that it would be a black ball?'

In getting some intuitive purchase on the similarity score used, it's helpful to note that the average similarity score of .966 corresponds to an average divergence of 20 percentage points. As can be gleaned from Table 1, in only one instance (round 9) did we see any substantial divergence, with the two markets closing at 72.32% and 24.32%, respectively, making for a final market price similarity score of .770 for that particular round, corresponding to a divergence of 48 percentage point. (Factoring out that particular round makes for an average similarity score of .988, corresponding to an average divergence of 11 percentage points.) However, when we turn to our final metric, we shall see that this divergence actually reflects well on the SRIM in question.

4.2.4. Comparative Accuracy

The final metric we looked at was accuracy, measured by way of the mean squared errors of the final market prices, compared to the actual distributions of black balls in the corresponding (virtual) urns from which the samples communicated to the participants were randomly drawn. The squared errors of the final market prices were as follows:

TABLE 2. SQUARED ERRORS

ROUND	MARKET TYPE	PRICE	DISTRIBUTION	SQUARED ERROR
1	Traditional	21.54%	1%	.042
	Self-resolving	22.93%		.048
2	Traditional	85.64%	81%	.002
	Self-resolving	94.85%		.019
3	Traditional	27.60%	22%	.003
	Self-resolving	9.97%		.015
4	Traditional	1.75%	7%	.003
	Self-resolving	9.51%		.001
5	Traditional	19.43%	14%	.003
	Self-resolving	2.12%		.014
6	Traditional	23.67%	39%	.024
	Self-resolving	30.49%		.007
7	Traditional	88.24%	66%	.050
	Self-resolving	74.86%		.008
8	Traditional	79.55%	75%	.002
	Self-resolving	70.37%		.002
9	Traditional	72.32%	28%	.196
	Self-resolving	24.32%		.001
10	Traditional	92.92%	90%	.001
	Self-resolving	97.74%		.006
	AVERAGE ERROR FOR TRADITIONAL MARKETS:			.033
	AVERAGE ERROR FOR SELF-RESOLVING MARKETS:			.012

Prices given for 'Yes' on the question 'If a ball were to be drawn from the urn, what's the probability that it would be a black ball?'; **distributions** correspond to the actual distribution of black balls in the (virtual) urn.

As can be seen from Table 2, the squared errors for the markets were relatively low, and moreover lower for the SRIMs than the TIMs. An average squared error of .012 for the SRIMs corresponds to an average error of 11 percentage points. By contrast, the average error of .033 on the TIMs corresponds to an average error of 18 percentage points. In other words, to the extent that the SRIMs deviated on average in their final market prices from those on the TIMs, the former were more accurate than the latter.[8] It is interesting to note, in particular, that in the case of the only two markets where there is a substantial divergence in final prices (i.e., round 9), it is the SRIM that is most accurate, landing at a final price of 24.32%, making for a squared distance of .001 from the actual distribution of 28%.

5. Two Potential Confounders

The values achieved on the above metrics are promising, as far as their consistency with the FVH is concerned. We therefore wanted to rule out that the similarity between the TIMs and the SRIMs across the experiment's ten rounds was a mere artefact of a confounding variable. Two potential confounders stand out:

First, we wanted to rule out that the participants were simply not clear on the difference between the two types of markets, and therefore exhibited similar trading behaviours on both for completely uninteresting reasons. In order to evaluate the likelihood of this being the case, we collected qualitative data on the participants' understanding of the difference between the two markets. We did this as part of a debrief phone call with each participant two days after the study (on July 21, 2017), under the guise of soliciting their feedback on the platform used. As part of the phone call, each participant was asked to explain how they were rewarded on the two types of markets, and was deemed to have a clear understanding of the difference between TIMs and SRIMs if they responded (without further prompting) that rewards on TIMs was a function of the actual distribution of black balls in the (virtual) urn, while rewards on SRIMs was simply a function of the market price at the time of closing. Four of the six participants showed a clear understanding of the difference between the two markets. One of those four even, unwittingly and unprompted, summed up the FVH to explain why their betting strategy on the SRIMs was identical to that on the TIMs, despite the fact that bets were rewarded in different ways on the two types of markets: it just wasn't clear to them what else they could possibly do but have their bets be informed by the samples received, and assume that the market price was a function of others continuously doing the same.

Second, as we saw in Section 4.1, the instructions for the SRIMs included the following: 'This [market] price represents the market's judgment on the probability that a black ball would be drawn, in light of the samples aggregated on the market through the bets placed by you and others.' Did this statement prompt the participants to take the question asked at face value? We have no way of ruling out that it did. If it did, that might mean that the similarity we saw across markets was at least partly due to the prompt making salient the possibility that others would take the question at face value (in which case, arguably, you should, too). At the same time, this would actually be good news, since it would suggest that getting people to take the question at face value might be quite easy—perhaps all you need to do is prompt them to do so, and they will. Moreover, *if* the prompt had an effect—and we

[8] In fact, if we compare the TIMs and SRIMs in terms of the average squared error for each second of the market from the point at which a bet has occurred on both markets—a more demanding accuracy measure than one framed solely in terms of the final market price, since it penalises markets for not quickly converging on the correct price—the SRIMs still come out more accurate, with an average squared error of .039 across all ten SRIMs, compared to .048 across all TIMs.

stress that we have no way of knowing whether it did—then that might just be because the idea of taking the question at face value comes fairly naturally to people, which, if anything, is congenial to the FVH. (We will return to this matter in the next and final section.)

In light of this, we are fairly confident that the similarity across the two types of markets in the study is not a mere artefact of the participants not being clear on the difference between TIMs and SRIMs, and that, to the extent that the similarity arose partly in response to the prompt offered, that would if anything be an interesting result in its own right, since it would suggest that having people bet in a manner that is consistent with the FVH is actually quite easy.

6. Limitations and Future Work

As mentioned at the outset, the aim of the study was two-fold: first, to remedy the complete absence of any publicly available, experimental data on SRIMs; and, second, to evaluate the FVH by determining whether trading behaviour can come out sufficiently similar across the two conditions, i.e., external resolution and self-resolution. The study shows that it's clearly possible to generate meaningful betting on SRIMs—betting that is moreover similar to what we see on otherwise identical TIMs.

As noted earlier, the sample of twenty markets was too small to enable us to determine whether SRIMs and TIMs will generally come out as similar as they did in this study. In order to determine that, we need to run a similarly designed study with a larger sample. Of course, what we will ultimately want to understand is *under what conditions* SRIMs can be expected to generate valuable outputs. It would make sense for future studies to look at one condition in particular: the presence of *market manipulation*. TIMs have showed a high degree of resilience in the face of manipulation attempts (Hanson and Oprea 2009; Hanson, Oprea, and Porter 2006; Oprea et al. 2007; Berg and Rietz 2014; Camerer 1998). But even if the FVH is correct, it might be that any convention on SRIMs to take the question at face value and bet accordingly will be undone by the slightest sign of market manipulation, which by definition involves trades or bets made in an attempt to move markets independently of the external facts referenced by the questions. Investigating the susceptibility of SRIMs to market manipulation will therefore be an important part of a future work on SRIMs.

Of course, even a large sample study of this kind will be subject to the same worries that affect all laboratory studies regarding whether the results will generalise to naturalistic settings. This worry has motivated several recent studies, regarding both the comparative performance of information markets (e.g., Buckley 2017) and their susceptibility to manipulation (e.g., Buckley and O'Brien 2015; Berg and Rietz 2014). For this reason, even if a future study finds support for FVH on account of finding that trading behaviour SRIMs and TIMs can be expected to come out sufficiently similar in experimental settings, it would make sense to then apply SRIMs in a non-laboratory environment, with non-stylised contract questions dealing with real-life decision problems. Crucially, such applications could not be *directly* comparative. Again, the very situations in which we would want to implement self-resolving markets are ones where the type of external resolution required for a TIM is not a viable option, as per what was argued in Section 2. Still, a more *indirect* form of comparison—say, with the type and level of volatility we typically see on TIMs, the degree of sensitivity to manipulation that we tend to see on TIMs, and so forth[9]—would be possible and worthwhile. Moreover, testing SRIMs in naturalistic settings will also be crucial when it comes to having the

[9] We can think of these and similar features as *structural* features of information markets, that can be evaluated independently of the substantive matters bet on.

relevant type of markets be accepted in supporting the decisions of actual practitioners, who might be uncomfortable with relying on support systems that have not been extensively tested in the field.

Another factor of relevance to commercial SRIMs is *question curation*. In particular, it would be worthwhile to test whether we will see a significant difference along aforementioned metrics when we vary the instructions given to the participants, and specifically how much stress is put on the difference between TIMs and SRIMs. As noted earlier, the instructions for the SRIMs in the study included the claim that the price 'represents the market's judgment on the probability that a black ball would be drawn, in light of the samples aggregated on the market through the bets placed by you and others.' Earlier, we raised the possibility that this might have prompted the participants to take the question bet on at face value. If that is so, it is possible that, the more salient the instructions make the difference between the two types of markets, the less similar trading behaviour on SRIMs and TIMs will be. This would be worth testing since any attempt to design commercial SRIMs will want to make sure it gets right the amount and level of detail of the instructions given to participants, for purposes of ensuring meaningful trading.

These are just some questions regarding SRIMs on which future work would do well to focus on. Of course, there are plenty of other aspects of SRIMs that we need to gain a better understanding of, in addition to those mentioned here. Hopefully, if nothing else, our study will help motivate others to investigate these and other aspects of a type of market that, while presently not well-understood, potentially constitutes a powerful alternative to more traditional information markets in cases where relying on these is not a feasible option.[10]

References

Abramowicz, M. 2007. *Predictocracy: Market Mechanisms for Public and Private Decision Making*. New Haven, CT: Yale University Press.

Ahlstrom-Vij, K. 2016. 'Information Markets.' In D. Coady, K. Lippert-Rasmussen, and K. Brownlee (eds), *The Blackwell Companion to Applied Philosophy*, Wiley-Blackwell.

Ahlstrom-Vij, K. ms. 'Coordination and Conflict on Self-Resolving Markets.' Unpublished manuscript available at http://www.ahlstromvij.com/blog/2016/10/17/srim.

Antweiler, W. 2012. 'Long-Term Prediction Markets.' *The Journal of Prediction Markets* 6(3): 43-61.

Berg, J. and Rietz, T. 2014. 'Market Design, Manipulation and Accuracy in Political Prediction Markets: Lessons from the Iowa Electronic Markets.' *Political Science and Politics* 47(2): 293–296.

Berg, J., Nelson, F., and Rietz, T. 2008. 'Prediction Market Accuracy in the Long Run.' *International Journal of Forecasting* 24: 285–300.

Buckley, P. 2017. 'Evidencing the Forecasting Performance of Predication Markets: An Empirical Comparative Study.' *The Journal of Prediction Markets* 11(2): 60-76.

Buckley, P., and O'Brien, F. 2015. 'The Effect of Malicious Manipulations on Prediction Market Accuracy.' *Information Systems Frontiers* 19(3): 611-623.

Camerer, C. 1998. 'Can Asset Markets Be Manipulated?' *Journal of Political Economy* 106: 457–482.

Chen, K.-Y. and Plott, C. 2002. 'Information Aggregation Mechanisms: Concept, Design and Implementation for a Sales Forecasting Problem.' CalTech Social Science Working Paper No. 1131.

Debnath, S., Pennock, D., Lawrence, S., and Giles, C.L. 2003. 'Information Incorporation in Online In-game Sports Betting Markets.' *Proceedings of the 4th Annual ACM Conference on Electronic Commerce* (EC'03): 258–259.

[10] The authors would like to thank Michael Abramowicz, Werner Antweiler, Patrick Buckley, and Robert Northcott for valuable comments on an earlier version of this paper, as well as Mike Halsall and Karl Mattingly at Dysrupt Lab for their support.

Deschamps, B. and Gergaud, O. 2007. 'Efficiency in Betting Markets: Evidence from English Football.' *The Journal of Prediction Markets* 1: 61–73.

Espinoza, N., Erdeniz, R., and Kolk, K (ms.), 'Risk Markets,' unpublished manuscript.

Forsythe, R., Frank, M., Krishnamurthy, V., and Ross, T. 1998. 'Markets as Predictors of Election Outcomes: Campaign Events and Judgment Bias in the 1993 UBC Election Stock Market.' *Canadian Public Policy* 24: 329–351.

Graefe, A., and Weinhardt, C. 2008. 'Long-term Forecasting with Prediction Markets—A Field Experiment on Applicability and Expert Confidence.' *The Journal of Prediction Markets* 2(2): 71-92.

Hahn, R. and Tetlock, P. 2006. *Information Markets: A New Way of Making Decisions*. Washington, DC: AEI Press.

Hanson, R. 2007. 'Logarithmic Market Scoring Rules for Modular Combinatorial Information Aggregation.' *Journal of Prediction Markets* 1: 3-15.

Hanson, R. 2013. 'Shall We Vote on Values, But Bet on Beliefs?' *Journal of Political Philosophy* 21(2): 151-178.

Hanson, R. and Oprea, R. 2009. 'Manipulators Increase Information Market Accuracy.' *Economica* 76(302): 304–314.

Hanson, R., Oprea, R., and Porter, D. 2006. 'Information Aggregation and Manipulation in an Experimental Market.' *Journal of Economic Behavior and Organization* 60: 449–459.

Horn, C. F., Ohneberg, M., Ivens, B. S., and Brem, A. 2014. 'Prediction Markets—A Literature Review 2014 Following Tziralis and Tatsiopoulos.' *The Journal of Prediction Markets* 8(2): 89-126.

Keynes, J. M. 2015. *The General Theory of Employment, Interest and Money*. In R. Skidelsky (ed.), *The Essential Keynes*, Penguin; originally published in 1936.

Klingert, F. M. A. 2017. 'The Structure of Prediction Market Research: Important Publications and Research Clusters.' *The Journal of Prediction Markets* 11(1): 51-65.

Lewis, D. 1969. *Convention: A Philosophical Study*. Cambridge, MA: Harvard University Press.

Luckner, S., Schröder, J., and Slamka, C. 2008. 'On the Forecast Accuracy of Sports Prediction Markets.' In *Negotiation, Auctions & Market Engineering, Lecture Notes in Business Information Processing* (LNBIP), edited by H. Gimpel, N.R. Jennings, G. Kersten, A. Okenfels, and C. Weinhardt, 227–234. Dordrecht: Springer.

Mattingly, K. and Ponsonby, A.-L. 2004. 'A Consideration of Group Work Processes in Modern Epidemiology.' *Annals of Epidemiology* 24(4): 319-323.

McHugh, P. and Jackson, A. 2012. 'Prediction Market Accuracy: The Impact of Size, Incentives, Context and Interpretation.' *The Journal of Prediction Markets* 6(2): 22-46.

McKenzie, J. 2013. 'Predicting Box Office with and Without Markets: Do Internet Users Know Anything?' *Information Economics & Policy* 25: 70-80.

O'Leary, D. E. 2011. 'Prediction Markets as a Forecasting Tool.' *Advances in Business and Management Forecasting* 8: 169-184.

Oprea, R., Porter, D., Hibbert, C., Hanson, R., and Tila, D. 2007. 'Can Manipulators Mislead Market Observers?' Chapman University, E.S.I. Working Papers 08-01.

Pennock, D., Lawrence, S., Nielsen, F.A., and Giles, C.L. 2001. 'Extracting Collective Probabilistic Forecasts from Web Games.' *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 174–183.

Polgreen, P., Nelson, F., Neumann, G., and Weinstein, R. 2007. 'Use of Prediction Markets to Forecast Infectious Disease Activity.' *Clinical Infectious Diseases* 44: 272–279.

Rajakovich, D. and Vladimirov, V. 2009. 'Prediction Markets as a Medical Forecasting Tool: Demand for Hospital Services.' *The Journal of Prediction Markets* 3: 78-106.

Rosenbloom, E.S. and Notz, W. 2006. 'Statistical Tests of Real-Money Versus Play-Money Prediction Markets.' *Electronic Markets* 16(1): 63-69.

Schelling, T. 1960. *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.

Servan-Schreiber, E., Wolfers, J., Pennock, D., and Galebach, B. 2004. 'Prediction Markets: Does Money Matter?' *Electronic Markets* 14(3): 243-251.

Spann, M. and Skiera, B. 2003. 'Internet-Based Virtual Stock Markets for Business Forecasting.' *Management Science* 49: 1310–1326.

Tziralis, G., and Tatsiopoulos, I. 2007. 'Prediction Markets: An Extended Literature Review.' *The Journal of Prediction Markets* 1: 75-91.