SELF-RESOLVING INFORMATION MARKETS: A COMPARATIVE STUDY

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

Abstract: Traditional information markets (TIMs) are resolved with reference to events external to the markets, such as some particular candidate winning an election. However, when making long-term forecasts or evaluating counterfactuals, such resolution is not an option. Hence, the need for *self-resolving* information markets (SRIMs), resolved with reference to features internal to the markets themselves. The present paper demonstrates experimentally that the *market profiles* of otherwise identical TIMs and SRIMs show significantly higher degrees of correlation than do randomly paired markets, and that the average *accuracies* of TIMs and SRIMs are practically equivalent. This supports the so-called *face-value hypothesis*, on which a convention will arise on SRIMs of taking the question under consideration at face value and betting accordingly, in the same way as on TIMs—in which case SRIMs have the potential of matching TIMs in accuracy while shedding their limitations in relation to long-term predictions and counterfactuals.

Key words: Information markets; Prediction markets; Self-resolving markets; Coordination games; Long-term forecasts

1. Why Self-Resolving Markets?

Information markets, sometimes referred to as prediction markets, are markets for placing bets on future or otherwise unknown events. On *traditional* information markets (TIMs), rewards are tied to events external to the market, such as a central bank raising or lowering the interest rate by some specific amount, or a particular candidate winning an election. The benefit of this type of resolution is fairly obvious: it creates clear incentives to bet in accordance with what one takes the relevant facts to be, and for those in the know to reveal their knowledge on the markets (Ahlstrom-Vij 2016). Consequently, it is no surprise that the price signals arising on such markets, if interpreted as probability assignments, generally have turned out to 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).[1]

At the same time, given this type of resolution mechanism, there is also 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 cases involving long time-horizons or counterfactual events, this is not an option. Consider long time-horizons first. Antweiler (2012) suggests that there are two main challenges for long-term information markets, understood as markets with a time-horizon measured not in weeks or months but in years. First, 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). Second, 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 problem about opportunity costs can be overcome by the market maker compensating traders through separate investments. But there is 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 be-

[1] See Tziralis and Tatsiopoulos (2007) for a comprehensive review of the literature on information markets covering the period of 1990 through to 2006, and Horn *et al*. (2014) for the period of 2007 until 2013.

tween 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 of Antweiler's challenges: that the market will suffer from low liquidity, owing to there being too big a discrepancy between the long time-horizon of the market and the short attention span of potential traders. One way of meeting this challenge 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 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 this 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—the 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. After all, one of the main attractions of information markets is exactly 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, the challenges associated with setting up a market that successfully predicts events far into the future very much remains. So, consider, next, the evaluation of counterfactuals. 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 exists. Consider an example: *Had Russia not interfered in the 2016 US Presidential election, Hillary Clinton would have won*. 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.[2] 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, 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. Still, as Hanson (2013) 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). However, by that same line of reasoning, 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.

Hence, the need for *self-resolving* information markets (SRIMs), resolved with reference to factors internal to the markets themselves. 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] 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

[2] Future (or indeed past) events relating to other elections and attempts at influencing them might of course offer *evidence* about whether the counterfactual statement in question is likely true or false, by making more or less likely claims about the underlying causal mechanisms. But such events cannot 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.

[3] As we shall see, the two exceptions here are Ahlstrom-Vij and Williams (2018) and Abramowicz (2007).

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, consider the fact that, on a TIM, 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 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—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 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 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—and speculative bubbles notwithstanding, trades are by convention often grounded in fundamentals. We typically 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.

It should be stressed that the suggestion is *not* 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 the main hypothesis of what's to follow. Because the hypothesis to be evaluated 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? 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*. 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 main 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.[4]

[4] The FVH's claim about saliency can be re-described in terms of what is sometimes referred to as *focal* or *Schelling points*, the latter being referred to in this manner on account of Schelling's (1960) work on the factors, locations, considerations, etc., that will for conventional reasons 'stand out' to people as particularly salient or otherwise relevant, and thereby be the ones on which people will tend to coordinate in coordination games. If we frame the FVH in these terms, then aforementioned worry about Keynesian beauty contests might be refor-

The FVH is significant because, if it is true, we can expect that, in the case of both TIMs and SRIMs, 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.

2. Putting the Face Value Hypothesis to the Test

How would we go about putting the FVH to the test? In line with a pilot study conducted prior to the large-sample study to be reported on below—a pilot study that suggested that TIMs and SRIMs *can* under fairly standard conditions come out highly similar, but was underpowered for purposes of saying anything more general than that (Ahlstrom-Vij and Williams 2018)—the strategy to be pursued in what follows is to identify a number of empirically testable ways (three, to be specific) in which TIMs and SRIMs might turn out to be similar, each of which would offer good evidence for the FVH in virtue of being explained by the FVH being true. For example, say that it turned out that the *price volatility* on SRIMs is more or less equivalent to that on yoked TIMs, where two markets are yoked in so far as they are for all practical purposes identical (more on this in Section 3), save for the manner in which bets are rewarded. That would offer evidence for the FVH, on account of how that similarity could be plausibly explained by participants on either market taking the question at face value, and betting accordingly. Hence, our first hypothesis:

H_{VOL}. Yoked TIMs and SRIMs exhibit practically equivalent degrees of volatility on average.

Taking inspiration from the common practice of measuring price volatility on a stock market in terms of standard deviation, mean degrees of volatility for SRIMs and TIMs will be defined as the mean standard deviation of the respective type of market. In the absence of any established norms on the matter, it was decided for purposes of testing whether those means are *practically equivalent* that the equivalency bounds were to be set in terms of +/-10 per cent of the mean volatility across both types of markets. In other words, if the relevant equivalency test (more on this in Section 4) suggests that the mean volatilities of SRIMs and TIMs do not diverge in either direction by more than 10 per cent of the pooled standard deviation, using a 95 per cent confidence interval, then they are practically equivalent for purposes of this hypothesis.

Furthermore, if it were to turn out that otherwise identical TIMs and SRIMs exhibit similar *market profiles*, that, too, would be good evidence for the FVH, given that participants' trading with reference to the external facts referenced in the question would go some way towards explaining that similarity. What would it mean for two markets to have *similar* marker profiles? The most straightforward way to operationalize such similarity is by way of the correlation between market prices. The challenge for formulating a testable hypothesis here is that there is not any *a priori* benchmark for what would constitute a *high* degree of correlation, representing a high degree of similarity. Clearly, we cannot expect a correlation of 1; but nor is there any particular degree of correlation of *less* than 1 that stands out as the obvious bar for a high correlation between markets. For this reason, the relevant hypothesis is best formulated in *comparative* terms. Specifically, if people trade with reference to the external facts referenced in the relevant question on SRIMs, we should expect that the correlation between yoked TIMs and SRIMs is significantly *higher* than for randomly paired (and as such likely *not* otherwise identical) TIMs and SRIMs. Hence, our second hypothesis:

---

mulated as follows: while the question bet on might constitute *a* Schelling point, it might not be the only or otherwise most prominent one. This is an instance of the more general worry about any process requiring coordination around a Schelling point. See, for example, Tabarrok (2018) for an interesting discussion on how that worry also crops up in relation to *Token Curated Registries*, which rely on mechanisms requiring that people coordinate on Schelling points relating to truth, quality, and the like. What we will see in relation to the information markets in this study, however, is that people in fact coordinating on the question asked offers the most plausible explanation of the results.

> H_{CORR}. The mean degree of correlation between market prices on yoked TIMs and SRIMs is significantly higher than the mean degree of correlation between an equally large set of randomly paired (i.e., non-yoked) TIMs and SRIMs.

After all, if the mean correlation between yoked pairs is significantly higher than between non-yoked pairs, that calls out for explanation, and one explanation is offered by the FVH: yoked pairs show a higher degree of correlation than do non-yoked pairs because the former, unlike the latter, are developing in response to participants' responses to the (same) external facts referenced in the question asked.

Finally, consider the fact that, if participants are betting with reference to the external facts on both SRIMs and TIMs, as per the FVH, we should expect such markets to exhibit similar degrees of *accuracy* with respect to their final market prices, whether or not they're highly similar in the path they take there. In light of that, our final hypothesis is the following:

> H_{FINAL}. The accuracy of final market prices between yoked TIMs and SRIMs will be practically equivalent on average.

The accuracy of market prices is for purposes of this study measured in terms of root-squared errors. Since an *error* measure, a lower score represents greater accuracy. Moreover, since a *root-squared* measure, it always takes on a positive value in the unit of the original scale, here: the unit interval. So, for example, if the final market price of a market is 0.7, and the correct price[5] is 0.8, then the accuracy of that market is 0.1. So, the *similarity* of accuracy will be measured in terms of the *difference* in root-squared errors of yoked markets. As in the case of information market volatility, here, too, there is no established norm for similarity. For that reason, it was decided that a 10-percentage point difference in either direction would serve as the benchmark for practical equivalency—any difference smaller than that can plausibly be considered practically insignificant.

3. Recruitment and design

1,000 unique participants were recruited from the online recruitment platform Prolific (www.prolific.ac). Existing studies suggest that Prolific offers greater diversity among participants, and a smaller proportion of 'professional survey takers' than other established platforms like Amazon Turk (Peer *et al*. 2017), as well as more control from the point of view of the researcher in terms of reliable pre-screening (Palan and Schitter 2018). Participants for this study were pre-screened to have at least A-level qualifications or equivalent (e.g., US high school) and an approval rating on the platform of at least 95 per cent. Participants were recruited using an online information sheet, introducing the study as follows:

> WHAT THE STUDY IS ABOUT
> Many predictions are off the mark because talk is cheap. Prediction markets[6], by contrast, force you to put your money where your mouth is by placing bets on future or otherwise unknown events. In this study, you'll get an opportunity to participate in a prediction market (using play rather than real money), together with other participants. You'll earn a base cash incentive for participating and, if you end up being the highest point scorer on the market, you also earn a bonus payment.
>
> WHAT TO EXPECT
> Once you join the study, you will be redirected to an online prediction market platform where you'll do the following:
> 1. Watch a 4-minute video, explaining what the study is about.
> 2. Do a quick comprehension check. This should take no more than 30 seconds.
> 3. Enter your Prolific ID.
> 4. Take part in the market. It will last for about 5-10 minutes.
> 5. Click a validation link, bringing you back to Prolific.

[5] On what constitutes *correct prices* for purposes of this study, see Section 3.

[6] Given that 'prediction markets' is the more established term for information markets, the study was framed in terms of the former in relation to the participants.

The video introduced and illustrated the notion of an information market, and the distinction between traditional, externally resolved information markets and self-resolving information markets. It explained that, on either market, the question participants would be asked was the following:

> Assume there's a large urn with black and white balls in it. Say that we were to draw a ball from it at random. What's the probability that it would be a black ball?[7]

The voiceover on the video elaborated as follows:

> If you think there are lots of black balls in it, you'll want to say that it's high. If you think there are only a few, you'll think it's low. How would you know? At the start of the market we'll draw one ball from the urn and let you know whether it's white or black. We'll draw more such samples (with replacement) at regular intervals. So, make sure to keep an eye out on the right [of the market interface] for those when you're on the market, as they'll be coming in frequently. Moreover, on the left [of the interface] you'll see a price graph that will develop in real-time as people share their estimates. In that respect, you might take the price signal at any given time as potentially telling you something about the samples others have received.

The markets used a standard version of Hansons's (2007) market scoring rule, sequentially rewarding participants to the extent that they, through their bets, moved the price signals towards the correct value, less any reward that is due to past participants moving the market in the same direction. This was explained to participants in the video. The comprehension check that followed the video asked the following, with the bracketed text varying depending on what type of market they were about to enter:

> You have been allocated to a [traditional/self-resolving] information market.
>
> That means that your bets will be rewarded with reference to
>
> - the actual proportion of black balls in the urn.
> - the market price at the time of closing.
>
> Please choose the option that applies and click 'Next'.

At that point, participants entered the custom built online information market platform. The user interface featured a graph on the left, developing in real time as participants placed bets on the market, and an 'inbox' on the right, where the participants received samples drawn randomly (with replacement) from the underlying distribution. Each participant started with 1,000 points to bet with. Underneath the graph, participants could click 'more likely' or 'less likely,' to bet a small number of points to nudge the market price either up or down.

200 markets, half of them TIMs and the other half SRIMs, were set up on the platform, each with five participants. Each TIM was yoked with a SRIM, such that there were 100 pairs of markets on the platform. Yoked markets were identical with respect to market duration (randomly assigned from the range of 240 to 360 seconds), the distribution of black balls in the (virtual) urn (in the range of 0 to 100 per cent), and the frequency with which random samples were sent to participants (ranging from every 15 to 30 seconds). For each market, the five participants received different samples. The samples were pre-generated, so as to ensure that participants on yoked markets received the same samples, with participant 1 on one market receiving the same samples as participant 1 on the corresponding, yoked market, and the same for participants 2 through 5. In other words, the only relevant difference between yoked markets was their manner of resolution.

The markets were run in the period of October 1 to November 10. Each participant was paid a base incentive of £1.30 for participating, amounting to an expected £7.8 per hour, given that the study took about 10 minutes to complete. Additionally, the highest point scorer on each market earned a £1

[7] This prompt might be taken to be problematic in light of the claim in Section 1 that we cannot evaluate counterfactuals by way of TIMs. But note that what the question asks about is a probability, which—as will become clear—in turn is interpreted as an actual (not counterfactual) proportion.

bonus, for a total of £2.30, in order to motivate participants to do well on the markets, and to create an incentive against destructive betting (where people bet with no regard for considerations about potential reward).[8] It was made clear to participants that, since all of their bets were to be made in play money, under no circumstances could they walk away from the study owing anything.[9]

4. Results

Section 2 outlined three empirically testable ways in which TIMs and SRIMs might turn out to be similar, each of which would offer good evidence for the FVH in virtue of being explained by the FVH being true. The following sections discuss whether the corresponding hypotheses were confirmed as part of the study.

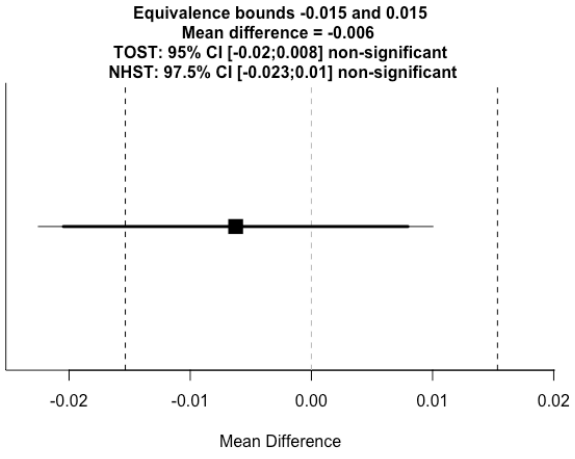
4.1. Volatility

Consider, first, our hypothesis about TIMs and SRIMs being practically equivalent with respect to their volatility:

H_{VOL}. Yoked TIMs and SRIMs exhibit practically equivalent degrees of volatility on average.

As mentioned earlier, the volatility of markets was measured by way of standard deviations. Measured thus, the mean volatility of TIMs and SRIMs was indeed similar, with a mean volatility of 0.150 for TIMs and 0.157 for SRIMs. For purposes of determining whether these values are practically equivalent, as per H_{VOL}, two one-sided *t*-tests were performed using *R*'s TOSTER package (*R* Core Team 2017; Lakens 2017), following Shapiro tests for normality and a Bartlett test for equality of variance.[10] For the two one-sided tests, an alpha of 0.025 was used for each, to get an alpha for the equivalency test of 0.05, and a corresponding 95 per cent confidence interval. As noted earlier, it was decided at the time of planning the study to use 10 per cent of the mean variance across TIMs and SRIMs (0.154) as the equivalency bound, thereby counting a mean volatility difference of +/- 0.015 as falling within the range of the practically equivalent.

As can be seen from Figure 1, TIMs and SRIMs did not come out statistically different on a standard null-hypothesis test, but nor did they come out practically equivalent within the bounds set:



[8] Looking ahead at the results regarding the accuracy of the markets in Section 4.4, there would seem to be no evidence of such destructive betting having taken place on any noticeable scale.

[9] Ethical approval for the study was obtained from Birkbeck, University of London, and documentation to this effect can be obtained upon request. In order to collect informed consent from each participant, the online information sheet included a consent note that made clear that, by clicking the link to be redirected to the platform, they were giving their consent to participate in the study, but that they were free to withdraw from the study at any point by simply leaving the platform. No participant requested to be withdrawn from the study. No other ethical concerns came up during the running of the study.

[10] The Shapiro tests and quantile-quantile plots suggested a non-normal distribution, but given the fairly large sample sizes (n = 100 for the TIMs and n = 100 for the SRIMs), this was disregarded for purposes of the subsequent *t*-tests. The Bartlett test suggested that the variance within the two samples was equal.

Figure 1. Equivalency test on mean difference in volatility between TIMs and SRIMs.

As such, the study failed to confirm H_{VOL}. While the mean volatility for the two types of markets came out similar, with a difference of only 0.006, they were on the bounds set prior to testing not practically equivalent.

4.2. Market Profiles

Turn, next, to our hypothesis about yoked TIMs and SRIMs being correlated as far as their market profiles are concerned:

> H_{CORR}. The mean degree of correlation between market prices on yoked TIMs and SRIMs is significantly higher than the mean degree of correlation between an equally large set of randomly paired (i.e., non-yoked) TIMs and SRIMs.

The mean Pearson correlation between yoked pairs was 0.166. To get a sense of what that degree of correlation might look like in practice, consider the following yoked markets with a correlation of 0.163:

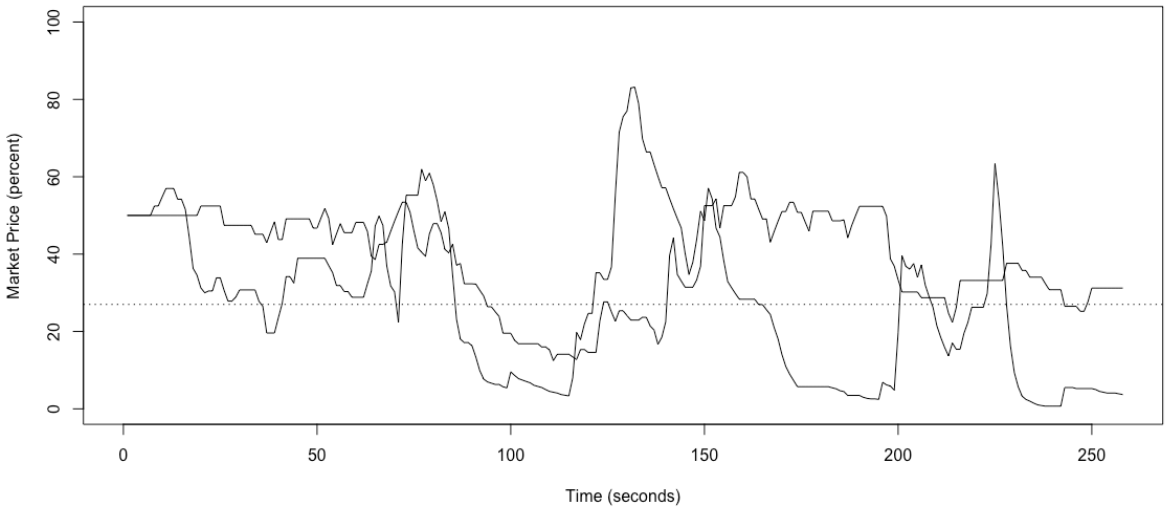


Figure 2. Two yoked markets with a correlation of 0.163, with the dotted line designating the true proportion of black balls in the (virtual) urn in question.

In order to evaluate H_{CORR}, the correlation between 100 randomly selected non-yoked pairs was calculated, making for a mean Pearson correlation of 0.020. A two-sample *t*-test suggested that the mean correlation of yoked pairs was indeed significantly higher than the mean correlation for non-yoked pairs (p = 0.00107; 95 per cent confidence interval: 0.059, 0.231).[11] As such, we have confirmation for H_{CORR}. In other words, otherwise identical TIMs and SRIMs (that is, markets that are identical in all other relevant respects than the manner in which the bets are rewarded) will tend to be significantly more correlated than TIMs and SRIMs that are not otherwise identical.

4.3. Accuracy

It was pointed out in Section 2 that, if participants on these markets are betting with reference to the external facts on both SRIMs and TIMs, as per the FVH, we should expect them to exhibit similar

[11] A Shapiro test for each sample, and a Bartlett test on both, suggested normal distributions and equality of variance, respectively. To protect against non-normality in calculating the correlations, the *t*-test was also carried out on the mean Kendall's τ correlations, with the same result: the mean correlations still came out significantly different (p-value = 0.001146).

degrees of accuracy with respect to their final market prices. Differently put, they should end up at around the same place, whether or not will tend to take the same routes there. Hence, our final hypothesis:

> H_{FINAL}. The accuracy of final market prices between yoked TIMs and SRIMs will be practically equivalent on average.

In evaluating H_{FINAL}, accuracy was—as noted at the outset—measured in terms of the root-squared error of final prices. The point of perfect accuracy (that is, an error of 0) was given by the true proportion of black balls in the (virtual) urn from which the samples distributed to participants on the corresponding market was drawn. As noted earlier, in terms of defining the bounds of practical equivalency when comparing the mean accuracy of markets, if it were to turn out that TIMs and SRIMs were on average less than 10 percentage points apart, then they can be considered practically equivalent in terms of their degree of accuracy.

Of course, before we consider any *similarity* with respect to the accuracy between TIMs and SRIMs, we are going to want to know whether they tend to be at all *accurate*. There is of course already a large literature demonstrating that TIMs will tend to be accurate with respect to their final market prices. The accuracy of TIMs in this study was in line with that literature: the average accuracy of TIMs came out at 0.183. As it happens, the SRIMs also came out highly accurate on average, with an average accuracy of 0.178. As such, the SRIMs were slightly more accurate on average than the TIMs (but as we shall see, not significantly so). More than that, as can be seen from Figure 3, the distribution of accuracy is heavily skewed towards the accurate end of the spectrum:

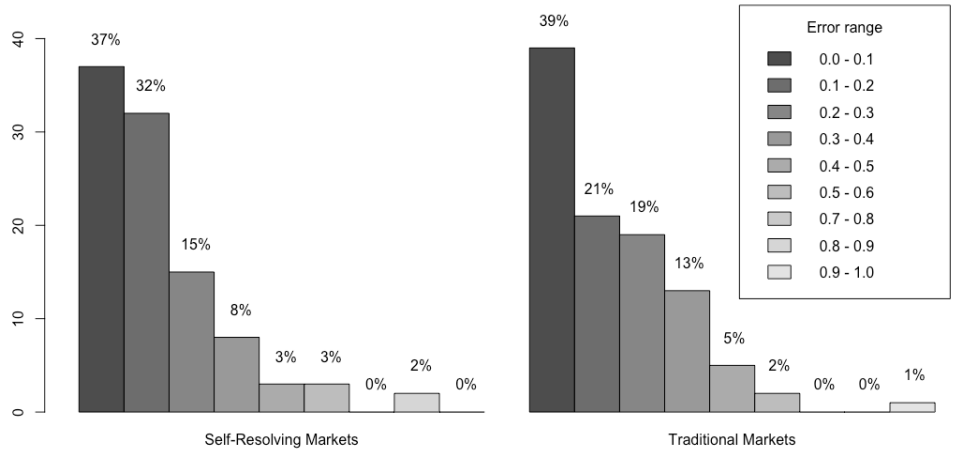


Figure 3. Proportion of markets the final prices of which fall within specific ranges of root-squared errors.

As we can see from Figure 3, 69 per cent of SRIMs and 60 per cent of TIMs exhibit a root-squared error of .2 or below, and very few markets of either kind exhibit high degrees of inaccuracy.

Moreover, an equivalency test was performed—again, using *R*'s TOSTER package[12]—which suggested that, not only were the degrees of accuracy of final market prices on TIMs and SRIMs not significantly different; additionally, they were practically equivalent within the +/- 0.1 bounds set earlier:

[12] As in the case of the equivalency test on comparative volatility, the Shapiro tests and quantile-quantile plots suggested a non-normal distribution, but this fact was disregarded given the fairly large sample sizes (n = 100 for the TIMs and n = 100 for the SRIMs). The Bartlett test suggested that the variance within the two samples was equal.

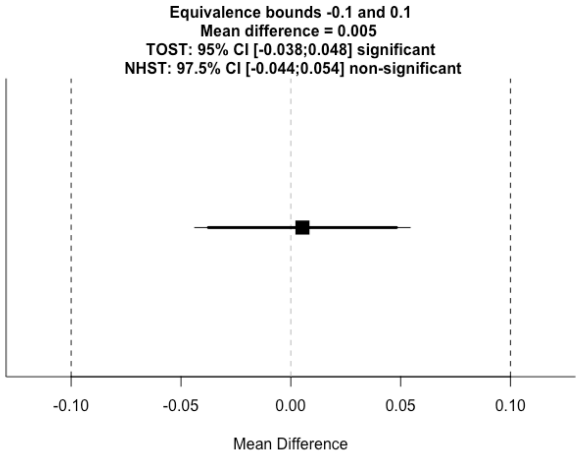


Figure 4. Equivalency test on mean root-squared error between TIMs and SRIMs.
(TOST p = 1.252954e-06 and 1.041178e-05)

As can be seen from Figure 4, the mean errors of TIMs and SRIMs would be equivalent all the way down to a difference of .05. Differently put, if TIMs tend on average to be accurate to degree x, in the range of 0 (perfect accuracy) to 1 (complete inaccuracy)—and judging by the accuracy of the markets run as part of this study, that degree is more likely to fall on the former end of that spectrum—one can expect that SRIMs on average will be accurate to degree x +/- .05, and *vice versa* for TIMs.

4.4. Comprehension

As noted earlier, each participant did a brief comprehension test before entering the market, to check whether they had likely understood the difference between traditional and self-resolving markets. This is important since any similarity across such markets would be uninteresting—and offer scant evidence for the FVH—if the participants involved behaved in similar ways across the two types of markets on account of not understanding the difference between the two.

Across the 200 markets, 71.4 per cent of participants gave a correct answer to the comprehension question. For those allocated to a TIM, 71.2 per cent of participants gave the correct answer; for those allocated to a SRIM, 71.6 per cent did. As such, a clear majority overall, and moreover an equally sized majority on the two types of markets, can be assumed to have understood the difference between TIMs and SRIMs. In light of this, it is unlikely that a lack of comprehension is driving any of the similarities identified above.

That said, it is worth noting that the practical equivalency in accuracy discussed in the previous section is robust all the way up to perfect comprehension. Specifically, an equivalency test on markets where *every* participant answered the comprehension correctly came out as follows[13]:

[13] A Shapiro test on each sample, and a Bartlett test on both, suggested a normal distribution and an equality of variance, respectively.

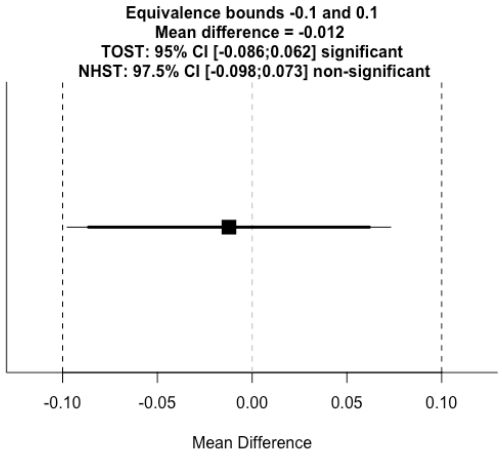


Figure 5. Equivalency test on mean root-squared error between TIMs and SRIMs with perfect comprehension (TOST p = 0.01068644 and 0.001907486).

By contrast, in relation to H_{CORR}, while the mean degree of correlation for yoked pairs where each participant answered the comprehension question correctly (0.312) came out higher than for non-yoked pairs (0.170), the difference was not significant. It is, however, quite likely that this result owes to the fact that there were only 12 such markets, or six pairs, meaning that the test in all likelihood was underpowered.

5. Discussion

Let us take stock. We have found support for two of the three hypotheses we formulated regarding potential dimensions of similarity between TIMs and SRIMs. More specifically, we did not find that the average volatility on TIMs and SRIMs were practically equivalent. However, we did find that the market profiles of yoked pairs were significantly *more* similar than the profiles of randomly selected, non-yoked pairs. We also found that the degree of accuracy of their final market prices were practically equivalent. Finally, we noted that the latter result was robust all the way up to perfect comprehension in regards to the participants' performance on the comprehension test.

At this point, it is worth reminding ourselves of our strategy for evaluating the FVH. It was suggested at the outset that the idea was to identify a number of empirically testable ways in which TIMs and SRIMs might turn out to be similar, each of which would offer good evidence for the FVH in virtue of being explained by the FVH being true. Looking at the hypotheses we have been able to confirm, it seems we now have such evidence. In particular, the fact that participants on SRIMs bet with reference to the (same) question asked on yoked TIMs and SRIMs would seem to offer a good explanation as for why the market profiles of such pairs show a significantly higher degree of correlation than do non-yoked pairs. And the similarity of TIMs and SRIMs with respect to accuracy offers even stronger evidence for the FVH. After all, it would seem very difficult to explain why the SRIMs came out as accurate as they did, and their degree of accuracy moreover was very similar on average to what we saw on the TIMs, if it were not for the fact that participants on the SRIMs took the question asked at face value, and bet with reference to the external facts on which the question asked turned. That is to say that, not only does the truth of the FVH serve to explain the results of the present study; additionally, it is difficult to imagine what would be a plausible, *alternative* explanation of these results.

So, in light of having been able to confirm both H_{CORR} and H_{FINAL}, there is good reason to believe that the FVH holds. This, in turn, means that SRIMs do indeed have the potential of incorporating the accuracy of TIMs, while shedding their particular weaknesses in relation to long-term forecasts and counterfactuals.

6. Directions for Future Work

Two broad areas of future work stand out as particularly relevant in light of the results above:

First, it would make sense for future studies to look at the effect on SRIMs 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 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.

Second, this study is 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, it would be a particularly natural next step to also apply SRIMs in a non-laboratory environment, with non-stylised contract questions dealing with real-life decision problems. Crucially, 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, meaning such situations would not be ones in which we would compare the performance of SRIMs against TIMs. Instead, implementation could proceed in two faces. In an initial phase, SRIMs would be implemented alongside established forecasting methods—be it deliberative groups of experts, or data scientific tools like machine learning on big data—to provide additional predictions for practitioners to consider in seeking a synthesis of available information across methods. In the next phase, once the experimental literature on SRIMs is larger and SRIMs have been implemented successfully (as judged by the relevant practitioners) along a range of forecasting tasks, our credence in their outputs might be sufficiently high to have them be implemented in a more free-standing manner, including in context where we might have such outputs trump that of alternative forecasting methods.

Naturally, 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, this study will help motivate others to investigate these and other aspects of a type of market that potentially constitutes a powerful alternative to more traditional information markets in cases where relying on these is not feasible.[14]

References

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

Ahlstrom-Vij, K. and Williams, N. 2018. 'Self-resolving Information Markets: An Experimental Case Study.' *The Journal of Prediction Markets* 12(2): 47-67.

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

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 Prediction 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.

[14] The study reported on in this paper was made possible through a grant from *the Research Innovation Fund* at Birkbeck College, University of London. The funders had no role in study design, data collection and analysis, the decision to publish, or the preparation of the manuscript. Many thanks to Michael Abramowicz, Ulrike Hahn, Mike Halsall, Robert Northcott, and Nick Williams for helpful discussions in relation to the study.

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

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.

Lakens, D. 2017. 'Equivalence Tests: A Practical Primer for *t* Tests, Correlations, and Meta-Analyses.' *Social Psychological and Personality Science* 8(4): 355-362.

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.

Palan, S., and Schitter, C. 2018. 'Prolific.ac—A Subject Pool for Online Experiments.' *Journal of Behavioral and Experimental Finance* 17: 22-27.

Peer, E., Samat, S., Brandimarte, L., and Acquisti, A. 2017. 'Beyond the Turk: Alternative Platforms for Crowdsourcing Behavioral Research.' *Journal of Experimental Social Psychology* 70: 153-163.

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.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

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.

Tabarrok, A. 2018. 'When Can Token Curated Registries Actually Work?' *Medium*; available at https://medium.com/wireline/when-can-token-curated-registries-actually-work-%C2%B9-2ad908653aaf.

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