No bang for your buck?

On accuracy incentives in expectation aggregation

December 8, 2022

Abstract

Due to the absence or cost of acquiring relevant historical data, many forecasting questions are answered by soliciting and aggregating the expectations of stakeholders. Prediction markets offer a particularly systematic way of doing just that, which moreover tends to yield accurate outputs. A common explanation for this is that such markets incentivise accuracy. However, the present paper reports on an experimental study involving participants being asked to perform two estimation tasks, mirroring the two main incentive structures on prediction markets, and finds minimal differences in accuracy on each task compared to a control condition. This calls into question the idea that prediction markets are successful on account of their incentive structure. Instead, it is argued that such markets are best understood as a special case of so-called expectation polls — polls or surveys asking people about participants' expectations (e.g., "Who will win the election?") as opposed to about their intentions (e.g., "How would you vote, if there were an election today?") — and that one plausible explanation for the relative success of prediction markets is at least partly that they ask the right type of question: a predictive question, tapping into respondents' expectations rather than their intentions or preferences.

Keywords: accuracy incentives; expectation aggregation; prediction markets; opinion polling

1 Introduction

Will some new piece of legislation negatively effect a specific, commercial sector? Will the COVID-19 pandemic significantly alter how companies manage offices? Will a new form of technology substantially change consumer behavior? Due to an absence of historical data, or

the cost of acquiring, processing, and maintaining such data, questions such as these are often answered by aggregating the judgments or expectations of different stakeholders (Hyndman 2018). The aggregation methods used in practice range from the less systematic (e.g., informal votes in a board room, or the ad hoc consultation of subject matter experts) to the more systematic. Prediction markets — markets for placing bets on future or otherwise unknown events — is one systematic method for expectation aggregation that moreover has been shown to offer a particularly efficient way of generating accurate estimates in a wide range of areas (Hahn and Tetlock 2006), including politics (Berg and Rietz 2014; Berg, Nelson, and Rietz 2008; Forsythe, Rietz, and Ross 1999), sports (Luckner, Schröder, and Slamka 2008; Deschamps and Gergaud 2007; Debnath et al. 2003), business (O'Leary 2011; Plott and Chen 2002; Spann and Skiera 2003), medicine and health care (Polgreen et al. 2007; Mattingly and Ponsonby 2014), and entertainment (McKenzie 2013; Pennock et al. 2001).

Given their success, we will want to know why they tend to generate accurate outputs. A common explanation is that prediction markets *incentivise accuracy*. As Sunstein (2006b) notes, "a correct answer is rewarded and an incorrect one punished. Hence, investors have a strong incentive to be right" (88). Specifically, as suggested by Hall (2010), there is "a powerful financial incentive for truthful revelation — a profit motive — that serves as a countervailing force to emotional, political, and professional motivations" (32). However, as suggested by Kenneth Arrow and others, it is also thought that "the potential for profit (and loss) creates strong incentives to search for better information" (Arrow et al. 2008: 877). Indeed, the idea that incentives on prediction markets perform this dual role of truthful revelation and information discovery is very common in the literature. Three representative statements are as follows:

> The price mechanism rewards participants for making accurate predictions (i.e., they win money) and punishes them otherwise (i.e., they lose money). Participants thus have an incentive to actively look for information and immediately and truthfully reveal it to the market (Graefe 2017: 38).

> [...] because people stand to gain or lose from their investments, they have a strong incentive to use (and in that sense to disclose) whatever private information they hold; they can capture, rather than give to others, the benefits of disclosure. [...] Prediction markets also impose strong incentives for traders to ferret out accurate

information (Sunstein 2006a: 205-6).

> [...] the markets provide an incentive to generate, gather and process information across information sources and in a variety of ways. Traders who perform these tasks well, prosper. Those who don't may go broke, may drop out of the market, and appear less likely to set forecast determining prices. (Berg, Nelson, and Rietz 2008: 286).

There are, however, a number of reasons to be skeptical of this explanation of prediction market accuracy. For one thing, as for *financial* incentives specifically, the accuracy difference between play- and real-money markets turns out to be either non-existent (Servan-Schreiber et al. 2004) or small and context dependent (Mchugh and Jackson 2012). For another, there is reason to believe that accuracy incentives of *any* kind should not make a difference to accuracy, at least in interesting cases involving non-trivial judgment tasks. As noted by Camerer and Hogarth (1999) in a review of the effects of financial incentives on performance in experimental settings, "incentives sometimes improve performance, but often don't" (34). Specifically, incentives are helpful primarily on clerical, and other clearly effort-responsive tasks. That is, if doing well simply means trying harder, then incentives can make a difference. By contrast, when doing well requires knowing more or having more skill, incentives do not help. Trying harder — whether to report your beliefs truthfully, or to collect further information — is of little use if you do not know what you are doing.

It is important to note that none of this impugns the practice of using incentives to get participants to engage with the relevant tasks to begin with. To avoid confusion, we need to distinguish between incentivisation for engagement and for accuracy. When *everyone* is rewarded (whether financially or not) for performing some particular task, we are incentivising for engagement; when participants (additionally) are rewarded *differently* depending on their level of performance (the better you perform, the more you get), we are incentivising for (in this case) accuracy. Throughout, when talking about incentivisation, it is the latter that is being referred to.

We have some reason, then, to suspect that accuracy incentives make no difference on, and as such do not explain the accuracy of, the type of expectation aggregation taking place on prediction markets. In order to determine whether this suspicion is borne out by the evidence, the present paper reports on an experimental study involving participants being asked to per-

form two estimation tasks. Specifically, to mirror the type of incentive structure that exists on prediction markets — so called external resolution versus self-resolution (Ahlstrom-Vij 2019) — participants in two treatment conditions were rewarded either with reference to successfully estimating some external fact or event (a form of external resolution), or with reference to successfully predicting the mean response of participants about these external facts and events (a form of self-resolution, since the question is "resolved" with reference to only internal factors). The results indicate significantly higher effort on the part of participants in the incentivised treatment conditions, as measured by completion time, yet a minimal differences in accuracy between each of the two treatment conditions and a control condition, where participants did not receive any accuracy incentive. This calls into question the idea that prediction markets are successful on account of their incentive structure.

In the discussion section, it is suggested that, in light of these results, prediction markets are best understood as a special case of what is sometimes referred to as *expectation polls*: polls or surveys asking people about participants' expectations in general (e.g., "Who will win the election?") (Murr, Stegmaier, and Lewis-Beck 2021; Rothschild and Wolfers 2012) or in relation to their social circles specifically (Ahlstrom-Vij 2022; Galesic et al. 2018) as, opposed to about participants' intentions (e.g., "How would you vote, if there were an election today?"). Such expectation polls tend to outperform traditional intention aggregation polls, likely because the former tap into people's non-trivial amounts of knowledge about the intentions and preferences of others, who are thereby implicitly sampled as well (Ahlstrom-Vij 2022). Given this, it is argued that one plausible explanation for the accuracy of prediction markets, especially compared to traditional surveys, is therefore at least partly that they ask the right type of question: a predictive question about respondents' expectations rather than their own intentions or preferences.

2 Method and sample

In order to evaluate whether accuracy incentives matter for the accuracy of prediction markets, two estimation tasks were developed as part of an experimental design. The first task asked participants to estimate the probability of drawing a black ball out of a (virtual) urn of black and white balls, in light of a sample of 10 balls that had already been drawn, showing 7 out of 10 being black. In the second task, participants were asked to estimate what proportion of the

UK's population would come to have received a first shot of a COVID-19 vaccine on February 1, 2021, in light of official figures estimating it to be 4.9% on December 15, 2020.

2.1 Recruitment

Participants were recruited via the online recruitment platform Prolific (www.prolific.co).[1] 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. 2022, 2017), as well as more control from the point of view of the researcher in terms of reliable pre-screening (Palan and Schitter 2018). In this case, participants were pre-screened to be residing in the UK, and to have an approval rating on the platform of at least 99%. The recruitment commenced on January 24, 2021, with 30 participants in a "soft-launch" to ensure that there were no issues with the route between the recruitment platform and the survey experiment, and concluded the next day, on January 25, with a total of 1224 responses.

2.2 Experimental conditions and incentives

Participants were randomly allocated to one of three conditions.[2] In the *external resolution condition*, participants saw the following prompt as they were answering the first estimation question, about the the probability of drawing a black ball out of a (virtual) urn:

> *PLEASE NOTE: The "correct" answer will be determined by the actual proportion of black balls in the (virtual urn). For example, if there are 70% black balls, the correct answer is 70%. Everyone giving the correct answer will take part in a lottery, where one person will be awarded £10, on top of the incentive received for participating.*

The same prompt appeared again, *mutatis mutandis*, as participants were answering the second estimation question, regarding what proportion of the UK's population would be reported by the government on February 1, 2021 as having received (at least) one shot of a COVID-19 vaccine.

In the *self-resolution condition*, participants saw the following prompt as they were answering the first estimation question:

[1] Ethical approval was obtained from the College Ethics Committee at [removed for blind review], prior to recruitment, with full details available on request.

[2] See Section 8.3 in the Appendix for the texts used to recruit participants to the different conditions.

PLEASE NOTE: The "correct" answer will be determined by the mean response given by participants. For example, if participants answer 1%, 2% and 3%, the "correct" answer is 2%. Everyone giving the correct answer will take part in a lottery, where one person will be awarded £10, on top of the incentive received for participating.

The same prompt appeared, *mutatis mutandis*, as participants were answering the second estimation question.

In the *control condition*, participants were not told anything about being paid beyond the incentive received for participating.

2.3 Incentives

The base incentive received for participating across all conditions was £0.38, corresponding to an estimated £7.60 per hour, given that taking part in the study was expected to take about 3 minutes. The idea behind making the potential reward of £20 (£10 per question) in the treatment conditions substantially larger than the base incentive — 52 times higher, to be exact — was exactly to incentivise the type of truthful revelation and information discovery that, as we have seen, are taken by many to explain the accuracy of prediction markets.

Respondents, moreover, seem to have responded to the incentive in the way we would expect if they were trying harder, or at least spending more time on the task. In the control condition, the median completion time was 2 minutes and 7 seconds. By contrast, in the self-resolution condition the median completion time was significantly higher, at 3 minutes and 1 second (Wilcoxon rank sum test; $W = 30995.5$, $p < 0.0001$). In the external resolution condition, the median completion time was also significantly higher than in the control, at 3 minutes and 25 seconds (Wilcoxon rank sum test; $W = 34689.5$, $p = < 0.0001$).[3] In light of this, it seems clear that participants invested greater effort by spending more time on the survey — and in so doing, potentially also seeking out further information (at least in the vaccination task) — in the two incentivised conditions.[4]

[3]The time spent on the survey was unavailable for 9 respondents, as they had not input their Prolific ID correctly, and could as such not be matched up with the relevant entries in the completion time data.

[4]Might the greater amount of time spent in the two treatment conditions simply reflect participants spending (more) time making sense of the instructions? If that were so, we would expect that those who spent more time exhibited greater comprehension, as captured by the manipulation check (see next section). However, while there was a 5 second difference between the median time of the two groups (3 min. 12 sec. for those failing the comprehension check, and 3 min. 17 sec. for those passing it), the distributions of time spent were very similar, and the difference not statistically significant (Wilcoxon rank sum test; $W = 34772$, $p = 0.7125$).

2.4 Disqualifications and manipulation check

19 of the 1224 responses were from respondents who had managed to participate more than once. These responses were removed prior to analysis, alongside 9 responses with missing estimates, leaving 1196 responses.

In order to evaluate the effect of accuracy incentives, it was important that those taking part understood the relevant incentivisation structure. For that reason, a manipulation check was included at the very end of the study, as follows (with 1 and 2 randomized):

> *On the previous pages, we told you how we would determine the "correct" answers to the questions. How will these be determined?*
>
> *1. By the actual proportion of black balls in the urn for the first question, and by the numbers reported on the official UK Government website for data on Coronavirus on February 1 for the second question.*
>
> *2. By the mean answer given by participants in this study for each of the two questions.*

The manipulation check disqualified 7 participants in the external resolution condition, and 91 in the self-resolution condition, leaving a final sample of 1098 participants, distributed across measured demographics as in Table 2 in the Appendix. The table shows a fairly balanced demographic distribution across the conditions, which suggests that random allocation in recruitment was largely successful, and not affected by disqualification.

3 Hypotheses

Two hypotheses were formulated. First, given the results discussed in Section 1 about incentives primarily making a difference in clerical and as such effort-responsive tasks, the following was expected:

> LOWER ERROR. For each of the two incentivised conditions (i.e., external resolution and self-resolution), the average error will be significantly lower in the urn estimation task, compared to the control condition.

Error was measured in terms of the average absolute percentage point deviation from the correct value. For example, if a participant estimated the proportion to be 30% and the correct

answer was 10%, the error for that participant would be 20 percentage points, and the error for the condition as a whole the average error for all participants in that condition. For the urn task, the correct value was stipulated as 70% in the external resolution condition, the idea being that, having nothing else to go on, participants would judge that the small sample of 7 black balls out of 10 reflected the distribution in the urn from which those samples had been drawn. In the self-resolution condition, the "correct" estimate was identified with the mean estimate, as per the instructions given to participants. The mean estimate ended up being 64%.

However, since we are ultimately investigating whether accuracy incentives make a difference, we also need to look beyond classical null hypothesis testing, and consider practically relevant effect sizes. After all, with a large enough sample, any effect will come out significant, whether substantial enough to be practically relevant. For that reason, we need to settle on an effect size that is practically relevant and consider whether any effect is significant using equivalency testing (Daniël Lakens, Scheel, and Isager 2018). What is a big enough difference for practical purposes in our case? There is no established standard for the simple reason that what level of error is tolerable is highly context dependent. The same would, therefore, need to go for any *difference* in error between competing method, which is what we are concerned with here. In the absence of a received standard, it was decided that a +/- 5 percentage point difference would be deemed to fall within the bounds of the practically equivalent. It is ultimately up to the reader to decide whether they deem these equivalency bounds appropriate for their purposes.

Hence, the second hypothesis:

> PRACTICAL EQUIVALENCY. For each of the two incentivised conditions (i.e., external resolution and self-resolution), the average error will fall within +/- 5 percentage points of, and as such be practically equivalent to, the average error in the control condition.

For the vaccination estimation task, the correct estimate in the external resolution condition was given by the number of people in the UK who on February 1, 2021, was reported as having received (at least) one shot of a COVID-19 vaccine, divided by the most up to date estimate of the UK population size, and rounded to the nearest whole number. On February 1, this corresponded to 9,296,367 / 66,796,807 = 13.92%, for a rounded value of 14%.[5] In the self-

[5]Vaccination figures were retrieved from the official UK Government figures at https://coronavirus.data.gov.uk/details/vaccinations, and UK population estimate from the UK Office for National Statistics at https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates, both on February 1, 2021.

resolution condition, the "correct" estimate was given by the mean estimate, which ended up being 15%.

4 Results

The mean estimate and error by condition and estimation task is given in Table 1, and the distribution of estimates in Fig. 1.

Table 1: Mean estimates (percent) and errors (percentage points) by condition (SD in paranthesis)

	External resolution	Self-resolution	Control
Urn: Mean estimate	67.32 (9.88)	63.58 (14.08)	66.33 (10.18)
Urn: Mean error	3.78 (9.51)	6.83 (13.88)	4.66 (9.76)
Vaccination: Mean estimate	17.41 (13.07)	14.98 (9.29)	15.92 (11.72)
Vaccination: Mean error	6.28 (11.98)	4.6 (8.14)	5.56 (10.51)

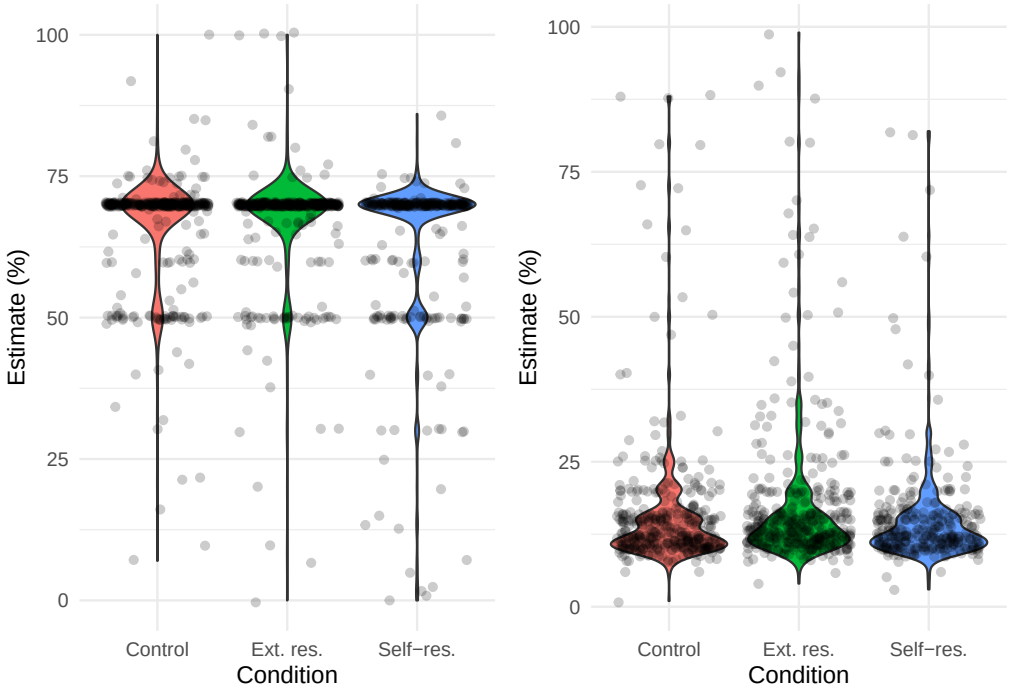

Figure 1: Distribution of estimates

Starting with the distribution of estimates in Fig. 1, the first thing to note is similarity across the three conditions. The second thing to note is that, unlike the vaccination estimation task, the distribution for the urn estimation task has not one but two clusters of estimates:

one around 70%, and a smaller one around 50%. It's clear that at least some users approached the estimate task against the background of a principle of indifference, likely considering that the small sample they had access to didn't settle the question of the correct proportion, which therefore should be assumed to be 50%. This was the case for a minority of respondents, and moreover a similarly sized minority across the conditions (7% for the external resolution condition; 11% for the self-resolution condition; and 10% for the control condition). Given this parity, the presence of such minority clusters do not affect any of our subsequent analysis about the comparative accuracy between conditions.

Turning to the errors, we see that the mean error is lower in the external resolution condition than in the control condition for the urn estimation task, and the error in the self-resolution condition is higher than in the control. The latter fact already tells us that LOWER ERROR — the hypothesis that, for *each* of the two incentivised conditions, the average error will be significantly lower in the urn estimation task, compared to the control — is not borne out by the evidence. For this reason, we reject LOWER ERROR.

However, we noted above that classical null hypothesis testing will not answer the question we are ultimately interested in, namely whether accuracy incentives make a *difference*. Specifically, we want to know whether PRACTICAL EQUIVALENCY holds — the hypothesis that, for each of the two incentivised conditions, the average error will fall within +/- 5 percentage points of, and as such be practically equivalent to, the average error in the control condition. For purposes of testing this hypothesis, the R package `TOSTER` (Daniel Lakens 2017; R Core Team 2018) was used to perform two one-sided t-tests in evaluating whether an effect size is large enough to be considered worthwhile.[6]

[6] As generating one confidence interval in this case requires performing two one-sided t-tests, an alpha level of 0.025 was used, to give a 95% confidence interval. To provide further protection against inflated Type I errors from performing several t-tests, the p-values for the euivalency tests were also adjusted using the Holm method. Adjusted p-values were < 0.0001 and < 0.0001 for the external condition in the urn task and < 0.0001 and < 0.0001 for the vaccination task; and < 0.0001 and 0.0014 for the self-resolving condition in the urn task and < 0.0001 and < 0.0001 in the vaccination task.

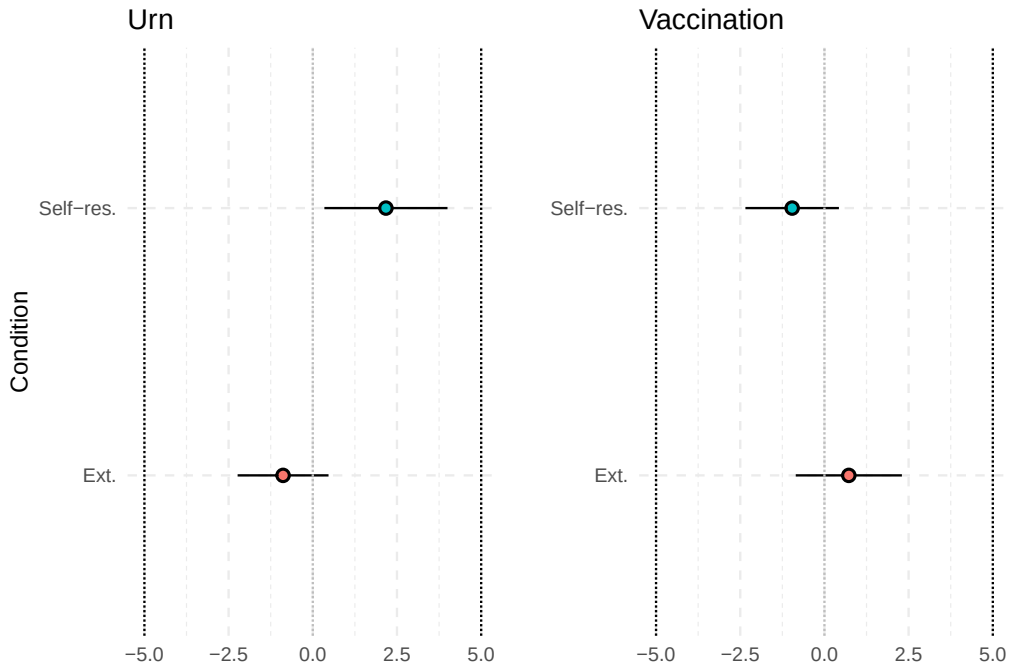

Figure 2: Equivalency tests on difference in mean error with a bound of +/-5 percentage points

If we look at any lack of overlap between the confidence intervals and the dashed vertical line at 0 in Fig. 2, we see the outcome of a set of traditional null hypothesis tests. Specifically, we see that the only significant difference is between the self-resolving condition and the control for the urn task. (This was the aforementioned effect that was in the "wrong" direction, as far as our LOWER ERROR hypothesis was concerned.) The remaining differences are not significant. However, as we are performing an equivalency test, we want to consider whether the confidence intervals clear the outer equivalency bounds at +/- 5 percentage points. In so far as they do, the mean errors for the relevant condition and the control are practically equivalent. Indeed, we see that, in the vaccination condition, which is arguably the most realistic estimation task of the two, the mean errors are practically equivalent all the way down to 2.5 percentage points. Consequently, PRACTICAL EQUIVALENCY is consistent with the evidence.

5 A third role for accuracy inventives?

In Section 1, we noted that the literature generally points to incentives on prediction markets performing the dual role of encouraging both truthful revelation and information discovery. The evidence provided for PRACTICAL EQUIVALENCY in the previous section calls into question whether incentives really offer such a mechanism — if they did, then we would have expected

to see a significant accuracy effect from the two incentivisation conditions, which we did not (indeed, on one case, we saw a significant *inaccuracy* effect).

However, there is another manner in which incentives might be able to increase accuracy on prediction markets, and that is by drawing in knowledgeable people, motivated by the incentives, who might not have joined the markets in the absence of such incentives. One way to spell this out is with reference to work suggesting that a typical prediction market works by having a fairly large number of uninformed traders provide liquidity, while a small number of marginal traders opportunistically take positions opposite to those of the uninformed, and in virtue of their trading volume also end up driving the market (Forsythe, Rietz, and Ross 1999; see also Oliven and Rietz 2004). These marginal traders tend to trade higher-than-average sums and to be active on the market on a higher-than-average number of days. They also earn higher-than-average returns, and are less prone to cognitive biases (such as assuming that their views are shared by others, or being inappropriately optimistic when interpreting evidence that supports their preferred outcomes).

Tying this back to accuracy incentives, the idea would be that incentives attract exactly such marginal traders, who in turn will drive accuracy. Notice, however, that suggesting that marginal traders *in actual fact* drive accuracy on incentivised markets does not go to show that they, counterfactually, would *not* have done so in the absence of incentives. As the results on the non-existent to small difference between play- and real-money markets shows us (Mchugh and Jackson 2012; Servan-Schreiber et al. 2004), people potentially join prediction markets for a variety of reasons. So, to investigate the relevant counterfactual, we would ideally want experimental evidence about the causal question of whether accurate participants are (more) attracted by the presence of (accuracy) incentives, compared to no incentives.

That, of course, is exactly the type of experiment we have reported on in this paper. When recruiting for this study, we recruited for the incentivised conditions by noting the additional, accuracy dependent incentive (i.e., a maximum of £20, on top of the £0.38 received simply for participating). For the control condition, the recruitment text said no such thing; it simply mentioned the fixed money paid for participating. (See Section 8.3 in the Appendix for the full recruitment texts.) This, in effect, provides a between-subjects test of the hypothesis under consideration that, if accurate respondents turn up in greater number in the presence of incentives, then we should expect to see more of these — and, as a result, greater accuracy — in the incentivised condition. But again, the evidence provided for PRACTICAL EQUIVALENCY

suggests that this did not happen.

6 Discussion

The results of this study call into question the idea that prediction markets tend to generate accurate outputs on account of incentivising accuracy, whether by motivating people to provide more honest estimates, collect further information, or — as discussed in the previous section — attracting accurate traders to at all join the market. As noted in 2.2, participants took significantly longer to complete the survey in both incentivised conditions, compared to the control group, suggesting greater effort and potentially also attempts on the part of participants to locate relevant information (at least in relation to the vaccination task). However, in neither condition was the level of accuracy significantly higher than in the control condition — on the contrary, they were practically equivalent. This offers evidence that incentivising for accuracy does not make a practical difference to estimation errors in a set of estimation tasks mirroring the type of incentive structures typically found on such markets.

This naturally raises the question of how we then are to explain the accuracy of such markets, if not with reference to incentives. One possible explanation is that prediction markets are best understood as a special case of what is sometimes referred to as expectation polls: polls or surveys asking people about participants' expectations in general (e.g., "Who will win the election?") (Murr, Stegmaier, and Lewis-Beck 2021; Rothschild and Wolfers 2012) or in relation to their social circles specifically (Ahlstrom-Vij 2022; Galesic et al. 2018) as, opposed to about participants' intentions (e.g., "How would you vote, if there were an election today?"). Such expectation polls tend to outperform traditional intention aggregation polls, likely because the former tap into people's non-trivial amounts of knowledge about the intentions and preferences of others (Nisbett and Kunda 1985), who are thereby implicitly sampled as well (Ahlstrom-Vij 2022).

Tapping into such implicit samples has two benefits. First, it has the potential of mitigating the type of respondent selection issues that typically affect surveys (Weisberg 2018), including non-response and coverage bias (Ahlstrom-Vij 2022; Galesic et al. 2018). That is, even if some particular group of people is systematically not sampled or generally non-responsive if sampled, there is a non-zero probability that someone sampled has information about that group that thereby gets factored into their reported expectations. Second, rather than being forced to

make one big inference about a preference distribution of a population on the basis of sampled preferences, as on the traditional opinion poll paradigm, someone conducting an expectation poll relies on a large number of inferences made by individual respondents, the errors of which will then cancel out if randomly distributed.

If this explanation is correct, the accuracy of prediction markets has nothing to with accuracy incentives, and is at least in part due to their asking the right type of question: a predictive question about respondents' expectations rather than about their own intentions or preferences. The qualifier 'in part' is important, since the present study does not speak to the question of the relative merits of simply surveying expectations and aggregating by way of some summary statistic such as the mean, as in this study, as opposed to aggregating those expectations through a more sophisticated prediction market mechanism, whereby each estimate dynamically moves the price signal (interpreted as the group's aggregate estimate) over time. This, however, is a question that has already been addressed by others. For example, in a sample of 535 participants and 113 forecasting questions, Dana et al. (2019) found that aggregating using an prediction market mechanism outperformed a simple mean approach, in terms of minimizing estimate error. This suggests that levels of accuracy are likely underestimated in the present study, and that there is an added accuracy benefit from aggregating specifically through a market mechanism, over an above any benefit derived from opting for an expectation aggregation approach over a traditional intention aggregation approach.

7 References

Ahlstrom-Vij, Kristoffer. 2019. "Self-Resolving Information Markets: A Comparative Study." *Journal of Prediction Markets*, February. https://doi.org/10.5750/jpm.v13i1.1687.

———. 2022. "On the Robustness of Social-Circle Surveys: Respondent Selection Issues, Egocentrism, and Homophily." *Electoral Studies* 75: 102433. https://doi.org/https://doi.org/10.1016/j.electstud.2021.102433.

Arrow, Kenneth J., Robert Forsythe, Michael Gorham, Robert Hahn, Robin Hanson, John O. Ledyard, Saul Levmore, et al. 2008. "The Promise of Prediction Markets." *Science* 320 (5878): 877–78. https://doi.org/10.1126/science.1157679.

Berg, Joyce E., Forrest D. Nelson, and Thomas A. Rietz. 2008. "Prediction Market Accuracy in the Long Run." *International Journal of Forecasting* 24 (2): 285–300. https:

//EconPapers.repec.org/RePEc:eee:intfor:v:24:y:2008:i:2:p:285-300.

Berg, Joyce E., and Thomas A. Rietz. 2014. "Market Design, Manipulation, and Accuracy in Political Prediction Markets: Lessons from the Iowa Electronic Markets." *PS: Political Science and Politics* 47 (2): 293–96. https://doi.org/10.1017/S1049096514000043.

Camerer, Colin F., and Robin M. Hogarth. 1999. "The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework." *Journal of Risk and Uncertainty* 19 (1): 7–42. https://doi.org/10.1023/A:1007850605129.

Dana, Jason, Atanasov Pavel, Tetlock Philip, and Barbara Mellers. 2019. "Are Markets More Accurate Than Polls? The Surprising Informational Value of "Just Asking"." *Judgment & Decision Making* 14 (3): 135–47.

Debnath, S, D Pennock, S Lawrence, and CL Giles. 2003. "Information Incorporation in Online in-Game Sports Betting Markets." In *Proceedings of the 4th Annual ACM Conference on Electronic Commerce (EC'03)*, 258–59.

Deschamps, Bruno, and Olivier Gergaud. 2007. "Efficiency in Betting Markets: Evidence from English Football." *Journal of Prediction Markets* 1 (1): 61–73. https://EconPapers.repec.org/RePEc:buc:jpredm:v:1:y:2007:i:1:p:61-73.

Forsythe, Robert, Thomas A Rietz, and Thomas W Ross. 1999. "Wishes, Expectations and Actions: A Survey on Price Formation in Election Stock Markets." *Journal of Economic Behavior & Organization* 39 (1): 83–110. https://doi.org/https://doi.org/10.1016/S0167-2681(99)00027-X.

Galesic, M., Bruine de Bruin W., Dumas M., Kapteyn A., Darling J. E., and E. Meijer. 2018. "Asking about Social Circles Improves Election Predictions." *Nature Human Behaviour* 2 (3): 187–93. https://doi.org/10.1038/s41562-018-0302-y.

Graefe, Andreas. 2017. "Prediction Market Performance in the 2016 U.S. Presidential Election." *Foresight: The International Journal of Applied Forecasting*, no. 45: 38–42. https://ideas.repec.org/a/for/ijafaa/y2017i45p38-42.html.

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

Hall, Caitlin. 2010. "Prediction Markets: Issues and Applications." *Journal of Prediction Markets* 4 (1): 27–58. https://ideas.repec.org/a/buc/jpredm/v4y2010i1p27-58.html.

Hyndman, & Athanasopoulos, R. J. 2018. *Forecasting: Principles and Practice, 2nd Edition*. OTexts: Melbourne, Australia.

Lakens, Daniel. 2017. "Equivalence Tests: A Practical Primer for t-Tests, Correlations, and Meta-Analyses." *Social Psychological and Personality Science* 1: 1–8.

Lakens, Daniël, Anne M. Scheel, and Peder M. Isager. 2018. "Equivalence Testing for Psychological Research: A Tutorial." *Advances in Methods and Practices in Psychological Science* 1 (2): 259–69. https://doi.org/10.1177/2515245918770963.

Luckner, Stefan, Jan Schröder, and Christian Slamka. 2008. "On the Forecast Accuracy of Sports Prediction Markets." In *Negotiation, Auctions, and Market Engineering*, edited by Henner Gimpel, Nicholas R. Jennings, Gregory E. Kersten, Axel Ockenfels, and Christof Weinhardt, 227–34. Berlin, Heidelberg: Springer Berlin Heidelberg.

Mattingly, Karl, and Anne-Louise Ponsonby. 2014. "A Consideration of Group Work Processes in Modern Epidemiology." *Annals of Epidemiology* 24 (4): 319–23. https://doi.org/https://doi.org/10.1016/j.annepidem.2014.01.001.

Mchugh, Patrick, and Aaron Jackson. 2012. "Prediction Market Accuracy: The Impact of Size, Incentives, Context and Interpretation." *Journal of Prediction Markets* 6 (2): 22–46. https://EconPapers.repec.org/RePEc:buc:jpredm:v:6:y:2012:i:2:p:22-46.

McKenzie, Jordi. 2013. "Predicting Box Office with and Without Markets: Do Internet Users Know Anything?" *Information Economics and Policy* 25 (2): 70–80. https://doi.org/https://doi.org/10.1016/j.infoecopol.2013.05.001.

Murr, A., M. Stegmaier, and A. Lewis-Beck. 2021. "Vote Expectations Versus Vote Intentions: Rival Forecasting Strategies." *British Journal of Political Science* 51 (1).

Nisbett, RE, and Z Kunda. 1985. "Perception of Social Distributions." *Journal of Personality and Social Psychology* 48 (2): 297—311. https://doi.org/10.1037//0022-3514.48.2.297.

O'Leary, DE. 2011. "Prediction Markets as a Forecasting Tool." *Advances in Business and Management Forecasting* 8: 169–84.

Oliven, Kenneth, and Thomas A. Rietz. 2004. "Suckers Are Born but Markets Are Made: Individual Rationality, Arbitrage, and Market Efficiency on an Electronic Futures Market." *Management Science* 50 (3): 336–51. http://www.jstor.org/stable/30046071.

Palan, Stefan, and Christian Schitter. 2018. "Prolific.ac—a Subject Pool for Online Experiments." *Journal of Behavioral and Experimental Finance* 17: 22–27. https://doi.org/https://doi.org/10.1016/j.jbef.2017.12.004.

Peer, Eyal, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. "Beyond the

Turk: Alternative Platforms for Crowdsourcing Behavioral Research." *Journal of Experimental Social Psychology* 70: 153–63. https://doi.org/https://doi.org/10.1016/j.jesp.2017.01.006.

Peer, Eyal, David Rothschild, Andrew Gordon, Zak Evernden, and Ekaterina Damer. 2022. "Data Quality of Platforms and Panels for Online Behavioral Research." *Behavior Research Methods* 54 (4): 1643–62. https://doi.org/10.3758/s13428-021-01694-3.

Pennock, David M., Steve Lawrence, Finn Årup Nielsen, and C. Lee Giles. 2001. "Extracting Collective Probabilistic Forecasts from Web Games." In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 174–83. KDD '01. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/502512.502537.

Plott, Charles R., and Kay-Yut Chen. 2002. "Information Aggregation Mechanisms: Concept, Design and Implementation for a Sales Forecasting Problem." Social Science Working Paper, 1131. California Institute of Technology.

Polgreen, Philip M., Forrest D. Nelson, George R. Neumann, and Robert A. Weinstein. 2007. "Use of Prediction Markets to Forecast Infectious Disease Activity." *Clinical Infectious Diseases* 44 (2): 272–79. https://doi.org/10.1086/510427.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rothschild, David, and Justin Wolfers. 2012. "Forecasting Elections: Voter Intentions Versus Expectations." *Brookings Report, November 1*.

Servan-Schreiber, Emile, Justin Wolfers, David M. Pennock, and Brian Galebach. 2004. "Prediction Markets: Does Money Matter?" *Electronic Markets* 14 (3): 243–51. https://doi.org/10.1080/1019678042000245254.

Spann, Martin, and Bernd Skiera. 2003. "Internet-Based Virtual Stock Markets for Business Forecasting." *Management Science* 49 (10): 1310–26. https://doi.org/10.1287/mnsc.49.10.1310.17314.

Sunstein, Cass R. 2006a. "Deliberating Groups Versus Prediction Markets (or Hayek's Challenge to Habermas)." *Episteme* 3 (3): 192–213. https://doi.org/10.3366/epi.2006.3.3.192.

———. 2006b. "Deliberation and Information Markets." In *Information Markets: A New Way of Making Decisions*, edited by R Hahn and P Tetlock. AEI-Brookings Joint Center

for Regulatory Studies.

Weisberg, Herbert. P. 2018. "Total Survey Error." In *The Oxford Handbook of Polling and Survey Methods*, edited by Atkeson Lonna and R. Michael Alvarez. New York, NY: Oxford University Press.

8 Appendix

8.1 Data

The data and code underlying the present paper can be accessed at [removed for blind review].

8.2 Participants

Table 2: Number of participants per condition by category

	Control	External resolution	Self-resolution
Female	255	287	221
Male	117	99	84
Gender not disclosed	12	16	7
Age: 18-24	92	78	48
Age: 25-34	116	126	101
Age: 35-44	79	89	66
Age: 45-54	45	56	52
Age: 55-64	40	35	30
Age: Over 65	9	10	11
Age not disclosed	3	8	4
Education: GCSE	52	36	40
Education: A-level	102	111	63
Education: Undergraduate	156	178	134
Education: Postgraduate	71	76	73
Education: None	3	1	2
N	384	402	312

8.3 Recruitment texts

Participants for the external and self-resolution condition, were recruited via the following text:

> *In this study, we will ask you two questions that will involve making an estimate. We will also ask you about your level of education.*
>
> *We expect this to take about 3 minutes, and will pay you £0.38 (corresponding to £7.60 per hour) for participating.*
>
> *Additionally, everyone giving the correct answer will take part in a lottery, where one person will be awarded £10, on top of the incentive received for participating.*
>
> *This means that, if you answer both questions correctly, you stand a chance to win £20, on top of the £0.38 you receive for simply participating.*
>
> *Participation is voluntary. You may withdraw at any point, and ask to have any data associated with your participation deleted by contacting the lead researcher, [removed for blind review]. If you withdraw, you forfeit your incentive payment.*

Participants for the control condition, were recruited via the following text:

> *In this study, we will ask you two questions that will involve making an estimate. We will also ask you about your level of education.*
>
> *We expect this to take about 3 minutes, and will pay you £0.38 (corresponding to £7.60 per hour) for participating.*
>
> *Participation is voluntary. You may withdraw at any point, and ask to have any data associated with your participation deleted by contacting the lead researcher, [removed for blind review]. If you withdraw, you forfeit your incentive payment.*