Reflexivity of Predictions as a Statistical Bias

Timotej Cejka

University of Chicago

Abstract

It has been argued that reflexive predictions pose a methodological problem to the social sciences by jeopardizing objective theory testing. This problem apparently arises from a reflexive prediction's ability to modify the evidential import conferred by an observation upon a tested theory. As a classic example, reflexive predictions can lead to a spurious confirmation of a theory (Merton, 1949). In a recent article, Kopec has argued that spurious confirmations constitute just one of multiple undesirable consequences of reflexive predictions, which he argues include both overstating and understating an observation's evidential support as well as mistaking *counter*-evidence for evidence (Kopec, 2011). I agree that reflexive predictions may alter the evidential import of an observation; however, I will argue that this is not a genuinely *methodological* problem for social science but instead a *technical* one. Specifically, I will show that the reflexivity of a prediction is merely a type of statistical bias and, as such, can be dealt with using standard econometric methods. After developing a new definition of reflexive predictions, I will show that econometric methods can in principle eliminate reflexivity from the predictions used to test theories in social science.

1 Reflexive predictions of probabilistic events

Suppose that disseminating the prediction of a political candidate's victory in an election would be sufficient to cause the candidate's victory. Clearly then, using this particular prediction to test a theory in social science would lead to the theory's spurious confirmation because the predicted event would not have occurred in the absence of the disseminated prediction. Since predictions of this type can make social scientists misjudge the relationship between observations and tested theories, it has been argued that reflexive predictions pose a methodological threat to social science (Kopec, 2011).

To define what it takes for a prediction to be reflexive, I will first analyze the definition proposed by Romanos and appraise an objection raised against it by Kopec. I will then move on to demonstrating that Kopec's own revised definition is unpalatable.

According to Romanos, a prediction is reflexive if and only if:

> **R:** the formulation/dissemination style of the prediction is a causal factor relative to the prediction's coming out true or false (Romanos, 1973).

Here formulation style (F-style) denotes the prediction's syntactic structure and the dissemination style (D-style) its manner of transmission and/or reproduction (Romanos, 1973).

The term "causal factor" in Romanos' definition means that there is some well-accepted or likely theory T_1 according to which the production of the prediction's F/D style (call this production event e_1) "is sufficient to bring about the occurrence of [the predicted event]" (event e_2) (Romanos, 1973). However, the requirement of a well-accepted (or likely) theory conflates P's *reflexivity* with the *knowledge* of P's reflexivity, because a prediction can turn out to be reflexive even if no such theory T_1 is available (Kopec, 2011). What's more, even in cases where e_1 is

sufficient for e_2, e_2 might still occur as an effect of the mechanisms posited by the tested theory. To evade this problem of over-determination, Kopec proposes that Romanos' definition be interpreted as the following. A prediction is reflexive if and only if:

> **R':** the formulation/dissemination style of the prediction is sufficient to switch the truth value of the prediction from what it would be if not disseminated.

Nevertheless, Kopec argues that the requirement of sufficiency leads to an overly restrictive characterization of reflexive predictions, because it implies that any prediction whose truth value depends even slightly on chance cannot be reflexive (Kopec, 2011).

To use Kopec's example, suppose that presidential candidates Jones and Smith are running for election. Of the voters who intend to vote, Smith is favored by 9 more than Jones, and the country's 10 undecided voters flip a coin to decide for whom to vote. Social scientists use their theories to predict that Jones will win the election. Were they to disseminate this prediction, 18 new people who did not initially intend to vote would be instigated to vote for Jones. In the case where the prediction is not disseminated, the probability of Jones winning is less than 0.1% as he would only win if all of the 10 undecided voters flipped the coin in his favor. I denote this scenario as $\mathbb{P}(J| \sim D) < 0.1\%$, where J represents Jones winning and $\sim D$ represents the condition of *no dissemination*. In the case of dissemination, the probability of Jones winning is more than 99.9% because he would only lose if all of the 10 undecided voters flipped the coin in Smith's favor. Thus $\mathbb{P}(J|D) > 99.9\%$.

In this example, dissemination (in some particular F/D style) would not be sufficient to change the truth value of the prediction because there remains some probability (albeit less than 0.1%) that Jones will *not* win the election even after the prediction has been disseminated. As a result, definition **R'** classifies this prediction as non-reflexive even though its dissemination

would cause the probability of Jones' victory to change dramatically from extremely unlikely ($\mathbb{P}(J| \sim D) < 0.1\%$) to extremely likely ($\mathbb{P}(J|D) > 99.9\%$).

This example poses a problem for Romanos' definition for two reasons. First, any prediction which changes the probability of the predicted event this drastically seems to represent exactly the kind of phenomenon which should be incorporated under the definition of reflexive predictions. Secondly, most, if not all, predictions in the social sciences are in fact partly determined by chance and therefore ruled out as *not reflexive* by Romanos's definition.

As a result, Kopec suggests that we extend the property of reflexivity to predictions of events which are partly determined by chance, by using the following definition. A prediction is reflexive if and only if:

> **K:** the formulation/dissemination style of the prediction is sufficient to change the probability of the predicted event occurring from what it would be if not disseminated (Kopec, 2011).[1]

I will denote the probabilities of an event e occurring given that the prediction of e's occurrence is not disseminated by $\mathbb{P}(e| \sim D)$ and disseminated (in some particular F/D style) by $\mathbb{P}(e|D)$. According to Kopec's definition then, a prediction is reflexive if and only if $\mathbb{P}(e| \sim D) \neq$

[1]Kopec substitutes the notion of "F/D style" for "mode of dissemination"; however, since this modification is inconsequential for my analysis, it will be omitted. Note also that Kopec refers to the phenomena characterised by definition **K** as *"weakly" reflexive predictions*, and to the phenomena characterised by definition **R'** as *"strongly" reflexive predicions*. Since, in the extreme, changing a prediction's probability amounts to switching the prediction's truth value, strongly reflexive predictions are clearly a proper subset of weakly reflexive predictions.

$\mathbb{P}(e|D)$. The prediction that Jones will win the election is considered reflexive under definition **K** because the small probability of Jones' victory in the case of no dissemination, $\mathbb{P}(J| \sim D) < 0.1\%$, changes to a very high probability in the case of dissemination, $\mathbb{P}(J|D) > 99.9\%$. In other words,

$$0.1\% > \mathbb{P}(J| \sim D) \neq \mathbb{P}(J|D) > 99.9\%.$$

Kopec's definition also allows that reflexivity vary in degrees, depending on the size of the change in the two probabilities. In Kopec's election example, the probability of the predicted event changes dramatically, and so this particular prediction represents a remarkably *strong* case of reflexive predications.[2]

2 Measuring the degree of reflexivity

Kopec's characterization is, however, too broad. In fact, the set of phenomena incorporated under Kopec's definition of reflexive predictions consists of *all* social-scientific predictions.

The problem is that Kopec's definition does not require the change in the probability of the predicted event to be in any sense *significant*. Unless the change in probability is exactly zero, i.e., $\mathbb{P}(e|D) = \mathbb{P}(e| \sim D)$, a prediction is considered to be reflexive. Thus, even the smallest change in the probability of the predicted event caused by dissemination will automatically classify a prediction as reflexive. Yet the actions of social scientists involved in producing a

[2]To avoid confusion, I will reserve the terms "strong" and "weak" to describe the size of the change in the probability of the predicted event due to dissemination. Thus, a prediction can be strongly reflexive even if it fails to switch the prediction's truth value, so long as it changes the probability of a predicted event by a sufficiently large amount.

prediction in *any* given F/D style inevitably interact with at least some of the initial conditions which lead to the predicted event and thereby change the event's probability ever so slightly. Thus, unless we stipulate an additional requirement, the set of phenomena described by Kopec's definition of reflexive predictions will also include non-reflexive predictions.

Secondly, Kopec's definition gives us no obvious method for distinguishing between strong and weak cases of reflexivity, because the term "change" could refer to absolute, proportional, or even statistically significant change. In fact, after interpreting Kopec's "change" as *absolute* difference, Lowe has been able to show that Kopec's definition classifies some seemingly strong cases of reflexive predictions as only exceedingly weak ones (Lowe, 2018).

To understand Lowe's argument, consider this experiment in behavioral economics which tests the (strong) free-rider hypothesis, positing that "the average investment in a public good will be 0% of agents' endowments" (Marwell and Ames, 1981). We will call this hypothesized event e. The experimenters found that, contrary to the free-rider hypothesis, the subjects typically end up investing around 50% of their initial endowments – except for economics graduates, who tend to contribute only around 20%. We will assume that the prediction of e's occurrence has been disseminated to the group of economics graduates (quite plausibly at some point in their economics curriculum) but not to the "non-economists".

Notice that since the group of non-economists typically contributes 50% rather than 0% of their endowments, the probability of the hypothesized event e is practically 0%. The same holds for the group of economics graduates who on average contribute around 20%. It follows that both $\mathbb{P}(e| \sim D)$ and $\mathbb{P}(e|D)$ are practically 0% and so even though there is some change in probability of e due to dissemination, the absolute difference between these two probabilities is "marginal at best". Hence although the prediction of event e is reflexive according to definition **K**, it only qualifies as an exceedingly weak case of reflexive prediction (Lowe, 2018). However,

this contradicts our intuition that by changing the average contribution from 50% to 20%, the dissemination of the hypothesis (specifically, to economics students) made the hypothesized event *e* much more likely than it would have been otherwise. Although Kopec is not committed to treating “change" as absolute difference, his definition is prone to Lowe’s objection because it fails to rule out this interpretation.

In order to address these objections against Kopec’s definition of reflexive predictions, I suggest the following revision. A prediction is reflexive if and only if:

> **RP:** the formulation/dissemination style of the prediction is sufficient to cause a change in the probability of the predicted event from what it would be if not disseminated, which would be *statistically significant* in an appropriate statistical model.[3]

A change is statistically significant if it is unlikely to have occurred randomly. For example, a model shows an association between dissemination in some particular F/D style and the probability of some predicted event to be statistically significant at the 5% significance level if there is only a 5% chance that the estimate of this association does not represent a *true* underlying correlation but rather a *random* association between two unrelated events. Furthermore, the statistical model which shows the change in probabilities to be statistically significant has to be *appropriate* according to our best (social-)scientific theories. Since statistical significance is always defined within some model of a statistical population, my definition’s requirement of an appropriate statistical model rules out pathological cases in which clearly implausible models

[3]Of course, by “statistically significant" I mean statistically significant at the standard significance levels. The question of what constitutes a *sufficiently* statistically significant change will be addressed later in the paper.

are fit in order to show non-reflexive predictions to be reflexive. Notice also that my definition (**RP**) is immune to cases in which a reflexive prediction is formulated and/or disseminated but cannot be shown to lead to a statistically significant change in probabilities because a model determining this statistical significance is unavailable.

We can now distinguish weak cases of reflexive predictions from non-reflexive predictions by using the *statistical significance* criterion. That is, unless dissemination of a given prediction (in some particular F/D style) can in principle be shown to be a statistically significant causal factor of the probability of a predicted event, then the prediction won't be classified as reflexive. Thus, unlike definition **K**, my definition does not treat *all* social-scientific predictions as reflexive.

When an appropriate statistical model shows that a prediction's dissemination in some particular F/D style is a statistically significant causal factor of the predicted event, then the model's estimate of the size of this causal effect determines the degree of the prediction's reflexivity. This allows us to distinguish *strongly* reflexive predictions from the *weak* sort, and we now also have a clear metric for measuring the relevant change in probabilities due to formulation/dissemination – the one defined in the statistical model. A prediction's dissemination in some F/D style affects the probability of a predicted event in a specific way depending on the context (e.g., the nature of the prediction). The mathematical relationship describing how social agents react to particular predictions, called the *reaction function* by Herbert Simon, will clearly need to be specified in the appropriate statistical model (Simon, 1954).

To illustrate my definition using Lowe's example of the free-rider hypothesis, suppose that the probability of a total contribution of 0% of the initial endowments is 2% for economics students, i.e., $\mathbb{P}(e|D) = 0.02$, and 1% for "non-economists", i.e., $\mathbb{P}(e| \sim D) = 0.01$. Although the absolute difference $\mathbb{P}(e|D) - \mathbb{P}(e| \sim D) = 0.01$ is vanishingly small, the change from $\mathbb{P}(e|D)$ to $\mathbb{P}(e| \sim D)$ could still be statistically significant. In fact, an appropriate statistical model would certainly

show this change to be not only statistically significant but also of a large magnitude because the probability of event e' s occurrence has increased substantially (doubled). Thus, if we assign the status of reflexivity to predictions based on statistical significance, and the degree of reflexivity based on the size of the F/D style's effect on the predicted event, then teaching the free-rider hypothesis to economics students would be treated as a *strong* case of reflexive prediction.

In practice, determining whether the dissemination of a prediction (in some particular F/D style) is a statistically significant causal factor of the predicted event can be accomplished using standard methods from econometrics and casual inference. For example, social scientists can perform a randomized controlled trial in which only one of two groups is exposed to dissemination of a prediction in some F/D style. The experimenters would then measure the probability of the predicted event in the group which experienced the specific F/D style and compare it to the group which did not experience any dissemination. In this deliberately simple example, the statistical significance and magnitude of the effect of dissemination in this F/D style could then be measured by the following regression model:

$$y_i = \alpha + \beta' \mathbf{x}_i + \gamma D_i + e_i \tag{1}$$

where i indexes the observed individual, y_i is the dependent variable (e.g., the probability that a person contributes 0% of their endowment to a public good), x_i denotes control variables (such as age, race, etc.), D_i is an indicator equal to 1 if the individual has experienced dissemination in the given F/D style (e.g., if she is an economics graduate) and 0 otherwise, the Greek symbols are coefficients and e_i is unobserved error. Finally, the experimenters would determine the statistical significance and magnitude of the change in the probability of the predicted event by examining the estimated coefficient γ in the regression output.

3 Eliminating reflexivity

Measuring reflexivity gives social scientists insight into how agents react to their predictions, i.e., it informs them about the *reaction function*. In practice, discovering the reaction function might involve no use of econometrics whatsoever (e.g., if the reaction function is obvious) or it might necessitate the use of more sophisticated econometric techniques than in the simple example outlined above. In fact, there are probably contexts in which estimating the reaction function would be an extremely challenging estimation exercise. The key point is, however, that estimating the reaction function can only constitute a difficult *technical* problem – it is not a *methodological* problem at all.

Once the effect of dissemination (in a particular F/D style) on the predicted event has been estimated, social scientists can use this knowledge of the reaction function to account for the reflexivity in their prediction. In fact, some knowledge of the reaction function allows social scientists to disseminate a prediction such that the occurrence of the predicted event will confer exactly the correct evidential support on their tested theory (Simon, 1954). In other words, there exists an F/D style for any social-scientific prediction which will allow the prediction to be confirmed non-spuriously. Confronted with an F/D style which would make their prediction reflexive, social scientists can always use a different F/D style which evades reflexivity. The problem of reflexive predictions thus completely reduces to the problem of estimating whether a particular F/D style is reflexive (and to what extent) and then applying this information correctly when choosing an F/D style for dissemination.

Suppose that a given F/D style of a prediction is a statistically significant causal factor of the predicted event, which is used to test a theory. The “true" underlying relationship between the predicted event and the mechanism causing it is represented by equation (1). Suppose also

that the social scientists testing the theory are completely unaware of the reflexivity of their prediction, and so they estimate the restricted model in equation (2).

$$y_i = \alpha + \boldsymbol{\beta}'\mathbf{x}_i + e_i \tag{2}$$

This econometric technique would make their estimates suffer from a paradigmatic example of statistical bias – specifically, omitted variable bias. However, if a given prediction is likely to be reflexive, then social scientific standards themselves demand that appropriate measures be taken against this bias. The fact that *some* social scientists might unknowingly test their theories using statistically biased results fails to raise any methodological issues for social science, as such cases represent examples of invariably *inadequate* social-scientific practice. The threat of reflexive predictions should therefore not be considered a methodological problem for social science but rather a technical issue of dealing with yet another type of statistical bias.

One could object that my definition of reflexive predictions fails to draw a clear line between weakly reflexive and non-reflexive predictions because it is unclear which *level* of statistical significance should be considered as appropriate. For example, economists typically use levels of one, five, and ten percent as a cut-off for statistical significance.

However, it should be left as a task for the social scientist who is performing the statistical modeling to determine which level of statistical significance is appropriate, since the choice of significance level depends on the social scientific theory used in the model as well as the model's empirical underpinnings (e.g., the sample size, statistical power, type of causal research design). If the model uses a level of five percent for the coefficients of all the other explanatory covariates (i.e., the control variables), then the same level should also be applied to the coefficient of the

F/D style.

One could also ask why five percent rather than for example six percent is used within social science itself. If the commonly used levels in social science are largely arbitrary, then my definition of reflexive predictions would inherit this arbitrariness.

However, the fact that five percent is a common level of significance does not entail that it is in any sense privileged over other levels, such as the arbitrary level of 6.52%. The fact that some social scientists assign special importance to round levels of statistical significance is merely another example of inadequate social-scientific practice.

4 Conclusion

My definition of reflexive predictions preserves Kopec's inclusion of predictions whose dissemination in a particular F/D style changes a predicted event's probability but fails to affect it so much as to switch its truth value. Nonetheless, I have argued that a prediction fails to be reflexive simply by being disseminated in an F/D style which changes the predicted event's probability. Instead, the F/D style must be a large enough causal factor of the predicted event so as to produce a *statistically significant* change in an appropriate statistical model. This allows my characterization of reflexive predictions to evade two major difficulties faced by Kopec's definition. Whilst my definition's criterion of statistical significance keeps out non-reflexive predictions, the size of an F/D style's effect in an appropriate statistical model provides a clear metric for determining the degree of a prediction's reflexivity. In fact, my analysis shows that the reflexivity of a prediction amounts to a form of statistical bias which can be addressed by the use of common econometric techniques. Thus, what has been called the *methodological* or "evidential" problem of reflexive predictions is rather a *technical* issue of statistical bias which can in principle be

eliminated by proper statistical methods within social science.

References

Buck, R. C. 1963. "Reflexive predictions." *Philosophy of Science*, 30, 359–69.

Grunbaum, A. 1963. "Comments on professor roger buck's paper 'reflexive predictions'." *Philosophy of Science*, 30, 370.

Kopec, M. 2011. "A more fulfilling (and frustrating) take on reflexive predictions." *Philosophy of Science*, 78, 1249–59.

Lowe, C. 2018. "The significance of self-fulfilling science." *Philosophy of the Social Sciences*, 48(4), 343–63.

Marwell, G. and Ames, R. E. 1981. "Economists free ride, does anyone else?." *Journal of Public Economics*, 15, 295–310.

Merton, R. K. 1949. *Social Theory and Social Structure*. Free Press, (Enlarged edn.) New York.

Romanos, G. D. 1973. "Reflexive predictions." *Philosophy of Science*, 40, 97–109.

Simon, H. A. 1954. "Bandwagon and underdog effects and the possibility of election predictions." *Public Opinion Quarterly*, 18, 245–53.

Vetterling, M. K. 1976. "More on reflexive predictions." *Philosophy of Science*, 43, 278–82.