Prediction scoring of data-driven discoveries for reproducible research

Anna L. Smith[1], Tian Zheng[2], and Andrew Gelman[2]

[1] *University of Kentucky,* [2] *Columbia University*

Abstract:

Predictive modeling uncovers knowledge and insights regarding a hypothesized data generating mechanism (DGM). Results from different studies on a complex DGM, derived from different data sets, and using complicated models and algorithms, are hard to quantitatively compare due to random noise and statistical uncertainty in model results. This has been one of the main contributors to the *replication crisis* in the behavioral sciences.

The contribution of this paper is to apply prediction scoring to the problem of comparing two studies, such as can arise when evaluating replications or competing evidence.

We examine the role of predictive models in quantitatively assessing agreement between two datasets that are assumed to come from two distinct DGMs. We formalize a distance between the DGMs that is estimated using cross validation. We argue that the resulting prediction scores depend on the predictive models created by cross validation. In this sense, the prediction scores measure the distance between DGMs, along the dimension of the particular predictive model. Using human behavior data from experimental economics, we demonstrate that prediction scores can be used to evaluate preregistered hypotheses and provide insights comparing data from different populations and settings. We examine the asymptotic behavior of the prediction scores using simulated experimental data and demonstrate that leveraging competing predictive models can reveal important differences between underlying DGMs. Our proposed cross-validated prediction scores are capable of quantifying differences between unobserved data generating mechanisms and allow for the validation and assessment of results from complex models.

Keywords and phrases: cross validation, experimental social science, model assessment, preregistration, reproducibility.

1. Introduction

Many scientific advances begin with exploratory investigations of observed data, but much of scientific practice relies on confirmatory analyses that evaluate data against a scientific hypothesis, accounting for uncertainty (Tukey, 1972). In recent years, the scientific community has become increasingly more open to statisticians' increasingly vocal warnings about this heavy reliance on null hypothesis statistical tests (NHST), and their p-values, that constitute much of confirmatory analyses (e.g., Ziliak and McCloskey, 2008; Nuzzo, 2014; Wasserstein and Lazar, 2016; Jeske, 2019). As a result, we have witnessed increasing research interest in alternatives to p-values and, more broadly, in methodology that can comprehensively account for the nuances of complex data, complex models, and the increasingly complicated algorithms necessary to estimate these models. Motivated by this discussion, we consider the statistical problem of exploring and understanding the behavior of a data-driven discovery across *two* sets of observed data that are believed to have been generated from similar processes or models (e.g., a study and its replication, as in Pawel and Held, 2020, or a pilot study and realized experimental data).

A key motivating example is the evaluation of preregistered hypotheses (Humphreys, Sanchez de la Sierra and Van der Windt, 2013; Gelman, 2013), often called *prediction scoring*, a research planning strategy which arose out of the frustration with p-values. Preregistration requires that

[*] To appear in *Statistics and Computing.*

researchers make publicly recorded predictions, often based on prior or pilot data, for the scientific hypotheses that will be assessed once the data has been collected. This forces researchers to clearly differentiate between confirmatory and exploratory analyses, which ensures that p-values for the confirmatory analyses can be safely interpreted as intended. After data collection, the researchers are faced with a natural question: *How well do the preregistered predictions align with the observed data?* This question requires a method of *scoring* the predictions, in the face of the materialized observations and the noise that comes with them.

This process of preregistration presents a unique statistical challenge: when we preregister our hypotheses we are making assumptions about the underlying data generating mechanism (DGM; e.g., it behaves like it did in this previous study; or we suspect this or that parameter to have a positive or negative significant effect). Commonly, preregistered studies formulate hypotheses in the form of NHSTs (e.g., about regression coefficients) and predict whether or not the associated p-values are significant; the preregistered predictions are then evaluated on a purely binary scale: is the p-value significant or not? Our proposed prediction scoring approach *quantifies* the differences between preregistered predictions (which are often informed by pilot data) and the realized experimental data; and results in a natural way to visualize these differences. Our approach represents a unique statistical perspective on preregistration that has gone largely unaddressed. Furthermore, prediction scores measure differences between DGMs, such as in the preregistration setting but also in more general cases. For example, prediction scores can leverage competing models to make discoveries about the underlying DGMs, as in the simulated data examples investigated in Section 4.

2. Our approach

Returning to our motivating example of preregistration, consider the setting in which we have access to some pilot data or prior study that informs our preregistered hypotheses.[1] Then, at the conclusion of a preregistered experiment, we are faced with two data sets: τ, from the pilot study, and τ', the realized experimental data. We assume that each of these datasets is generated from some unobservable data generating process—F and F'—and we are interested in learning about underlying differences between F and F' (a definition for DGMs is discussed in Section 3). We propose evaluating the difference between (1) the set of (preregistered) predictions, $\hat{\tau}'$ and (2) the observed experimental data, τ', from F'. These predictions, $\hat{\tau}'$, are predictions for the experimental data, τ', and are obtained from $\hat{f}^{\tau}$, a model for F that is trained on the pilot data τ (See Figure 1). In this section, we discuss some motivating ideas for how to use this comparison, between $\hat{\tau}'$ and τ', to learn about the quantity of interest, the underlying distance between F and F'. This comparison needs to decompose the total error (i.e., the observed difference between these model-based predictions, $\hat{\tau}'$, and the observed experimental data, τ') into estimation error (from estimating F by $\hat{f}$) and model differences, i.e., true differences between the underlying DGMs.

Why NHSTs are often insufficient. To better understand the nuances of this type of comparison, consider an example with a simple linear regression model with data sampled from each DGM (as in the right panel of Figure 1). Common practice is to predict the outcome of NHSTs for parameters in an assumed predictive model. When we have access to pilot data, we could simply merge the datasets and construct indicator variables, allowing for varying intercepts or slopes across the two samples. Hypothesis tests for the corresponding regression coefficients can nicely summarize these specific types of differences between the two DGMs; see also Figure 7 for a full example with real data. But for more general cases, evaluating differences between DGMs is not straightforward.

[1] This is the case we will assume throughout the rest of the paper; if this is not the case, we might considering simulating some potential pilot data that incorporates whatever scientific beliefs we hold about the process and wish to study in our preregistered hypotheses

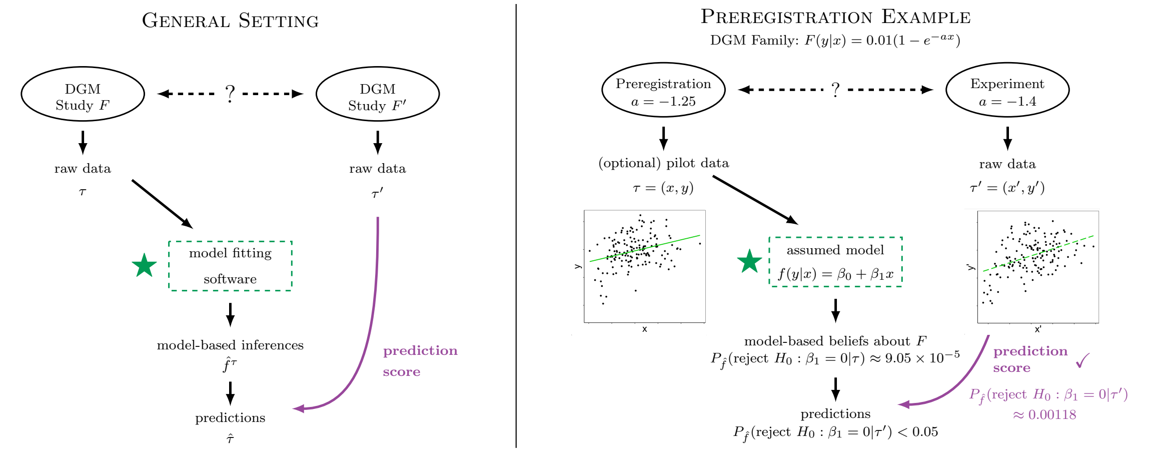
General Setting
DGM Study F
?
DGM Study F'
raw data τ
raw data τ'
model fitting software
model-based inferences $\hat{f}^{\tau}$
predictions $\hat{\tau}$
prediction score
Preregistration Example
DGM Family: $F(y|x) = 0.01(1 - e^{-ax})$
Preregistration $a = -1.25$
?
Experiment $a = -1.4$
(optional) pilot data $\tau = (x, y)$
raw data $\tau' = (x', y')$
y
x
y'
x'
assumed model $f(y|x) = \beta_0 + \beta_1 x$
model-based beliefs about F $P_{\hat{f}}(\text{reject } H_0 : \beta_1 = 0|\tau) \approx 9.05 \times 10^{-5}$
predictions $P_{\hat{f}}(\text{reject } H_0 : \beta_1 = 0|\tau') < 0.05$
prediction score ✓
$P_{\hat{f}}(\text{reject } H_0 : \beta_1 = 0|\tau')$ ≈ 0.00118

Fig 1. *Prediction scores are meant to measure the agreement between predictions and realized data. For preregistered studies (right panel), common practice is to specify predictions in the form of null hypothesis statistical tests about model parameters; these predictions are then scored on a binary scale. We formalize (and broaden) the concept of prediction scores as model-based continuous functions of predicted observables for the realized experimental data, τ', obtained from the assumed predictive model trained on some pilot or prior data, τ (left panel).*

For example, in the simple example in the right panel of Figure 1, the NHST from a linear regression is incapable of detecting the more nuanced true difference between F and F'. In practice, simple linear regression simply cannot sufficiently summarize the scientific process under consideration and, even when more complex statistical models can be used, relying on a single parameter or summary measure is typically unsatisfying.

Instead, we propose using appropriate predictive models to formulate scores for the predictions from one DGM F against observables from another, F'. As in predictive inference, this has the advantage of providing results that are highly interpretable and of direct interest to substantive researchers while also allowing for direct validation (in a way that is simply impossible for model parameters; Geisser, 2017; Billheimer, 2019). Further, this allows us to incorporate highly complex statistical models in exploring the unobserved DGMs, beyond linear regression models, and to move beyond effects described by a single parameter. As in posterior predictive checking for Bayesian models, this choice allows for great flexibility in the types of differences that can be uncovered between the underlying DGMs.

Controlling for model fit issues. In some ways, prediction scoring is similar to measuring predictive accuracy, but the goal here is different. Using the notation in Figure 1, both predictive accuracy and prediction scoring focus on discrepancies between $\hat{\tau}'$ and τ' but predictive accuracy uses these discrepancies to evaluate how well a model, $\hat{f}^{\tau}$, matches the true underlying DGM, F (and often which of among a set of competing $\hat{f}^{\tau}$s provides the best match). In this setting, it is typically assumed that $F = F'$; comparisons of this type are often called validation studies. In the prediction scoring setting, we do not assume that $F = F'$. As a result, typical predictive accuracy metrics are model-dependent measures that mix together the discrepancies between the modeling framework and the underlying DGM family (i.e., model fit issues) *and* any differences between the DGMs.

For this reason, we propose comparing validation summary measures to similar measures obtained from cross validation. The resulting prediction scores are then adjusted for cross validation's estimate of how well the model fits the data. Additionally, unlike other measures of model fit, cross validation is a general procedure that can accommodate many modeling frameworks (although appropriate partitioning can be difficult for dependent or hierarchical data; Racine, 2000; Gelman, 2006; Roberts et al., 2017) and is a clear analogue for the traditional validation procedure.

The role of the predictive model. In our approach, while the cross validation loss statistics help to normalize or account for some of the effects of a poor modeling choice, the chosen predictive model does impact the resulting prediction scores. As a result, the predictive model acts as a lens through which we can view differences between the two underlying DGMs, which themselves cannot be directly observed. Naturally, different models offer different perspectives and in the application described in Section 4 we leverage this aspect of our prediction scoring methodology by examining suites of non-nested predictive models to uncover distinct types of differences between the underlying DGMs.

Our proposed prediction scores. We provide full details of our proposed prediction scoring methodology in Section 3. In short, the idea is to learn about the distance between F and F' by comparing between model-based (preregistered) predictions, $\hat{\tau}'$, to observed experimental data, τ', using flexible loss functions to summarize the comparison and cross validation (along with subsampling of τ') to adjust for error due to model fit (see Figure 2; a more detailed discussion of this figure is given in Section 3). In Section 4, we extend our approach to consider differences across settings of a single simulated experimental setup and examine the probabilistic behavior of our proposed scores across many repetitions of the experiment. In Section 5, we return to the motivating setting of preregistration and demonstrate how prediction scoring can be used to evaluate preregistered predictions from a human behavior experiment. We discuss directions for future work in Section 6. Related methods for assessing predictive accuracy in the model selection setting and a review of recent advances in cross validation approaches are discussed in the appendix.

3. Cross-validated prediction scoring

In our approach, we will use the term *data generating mechanism* to refer to the unobservable underlying stochastic process that describes the scientific phenomenon under study. To be precise, we will conceptualize a DGM as a particular member of a family of probability distributions that represent a set of (model) assumptions about the scientific phenomenon

For any family of data generating mechanisms, we are interested in estimating a distance, or measure of dissimilarity, between different members of the same family.

Definition 3.1. For a particular family of data generating mechanisms, let the dissimilarity between any two members of the family be represented by

$$\Delta_{DGM} = d\left(F, F'\right)$$

where F and F' are both members of a particular DGM family and the choice of the dissimilarity function d is motivated by the form of the DGM family.

We will focus on regression-like settings where we are interested in learning about the relationship between a target variable, y, and a vector of inputs, X. As a result, we will think of the underlying DGM F as representing the joint population distribution for the data, but that research interests are focused on learning the conditional relationship, $p(y|X)$. This is a natural framework since, for example in a human behavior study, we may be interested in learning about the conditional relationship between individuals' demographic characteristics and some behavioral outcome but believe that a joint distribution describes the way that the sampled data are drawn from the population

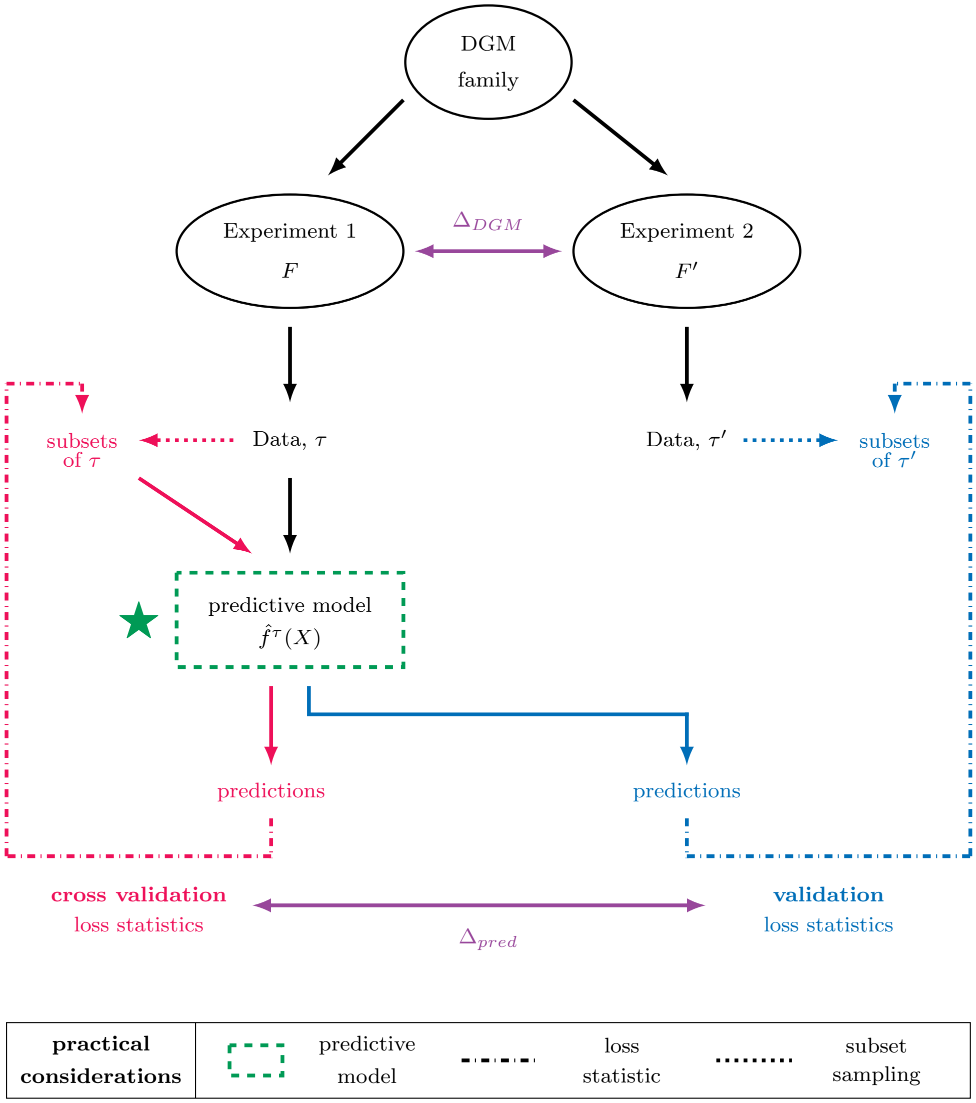

Fig 2. *General outline of the proposed prediction scoring methodology for generic data generating mechanisms, F and F'. We observe datasets $\tau = ((x_1, y_1), \dots (x_N, y_N))$ with data points drawn from F and $\tau' = \big((x_1', y_1'), \dots (x_N', y_N')\big)$ with data points drawn from F'. We believe that the target variable y is related to a vector of inputs X and estimate a predictive model, $\hat{f}^{\tau}(X)$, based on τ.*

of all possible studies (i.e., for a new sample, we would observe different individuals with different demographics and different behavioral outcomes).

3.1. Predictive accuracy and test error

To motivate the form of the prediction scores proposed here, we first review classical definitions of predictive error. As mentioned previously, the prediction scores essentially compare model-based predictions to real-world observations (see Figure 1), and so traditional model assessment tools are useful. In the model assessment setting, we are interested in evaluating the performance of a final selected model by estimating the model's predictive error *for new data*. Let $\hat{f}^{\tau}(X)$ be a predictive model estimated from a fixed dataset $\tau = ((x_1, y_1), \ldots, (x_N, y_N))$ drawn from F. For an appropriate loss function, $L(y, \hat{f}^{\tau}(X))$, model assessment tools typically estimate the conditional test error (also called the prediction or generalization error; Hastie, Tibshirani and Friedman, 2017), which depends on the particular (fixed) training set τ:

$$\mathrm{Err}_{F|\tau} = \mathrm{E}_{X,y\sim F}\left(L\left(y, \hat{f}^{\tau}(X)\right)|\tau\right).$$

This is the expected error for the predictive model trained on the dataset, τ, and the expectation is over all new data drawn from F but is conditional on observing the particular training data in τ. In practice, a new dataset is typically not available, and so this error is estimated using only the original observations, for example by performing cross validation (or bootstrapping or by calculating AIC, BIC, etc.) which estimates the expected test error, $\mathrm{E}_{\tau}[\mathrm{Err}_{F|\tau}]$, an average of the (conditional) test error over all possible training sets τ.

In the prediction scoring setting, we are considering two data generating mechanisms, or population distributions, F and F', each of which are joint distributions for the data. From each of these distributions, we observe a dataset, τ and τ' respectively. In the model assessment setting we care about the difference between F and $\hat{f}$ (i.e., how well does the model estimate the truth), whereas in the prediction scoring setting we care only about differences between F and F'. Since we cannot observe the DGMs directly, we can use a predictive model to summarize differences in the observed data; in this sense, the predictive model is like a nuisance parameter that we cannot avoid since the DGMs themselves cannot be directly observed (we only observe the datasets τ and τ').

Alternatively, we could consider measuring differences between the DGMs through explicit differences across the datasets themselves. For example, consider the test statistic for the two-sample t-test which is a function of the sample means. This test (and others like it) assume an underlying parametric model; the test is designed to detect differences between parameters from this model.

A natural model-based approach is to perform validation, where the predictive model is trained on (some subset of) the first (training) dataset, τ, but evaluated in the context of the new (test) data, τ'. The estimate of predictive error given by validation is typically of the following form,

$$\mathrm{Val}(\hat{f}^{\tau}) = \frac{1}{N'}\sum_{i=1}^{N'} L\left(y'_i, \hat{f}^{\tau}\left(x'_i\right)\right)$$

where N' is the number of observations in τ'. In this sense, validation error estimates a different conditional test error, given by

$$\mathrm{Err}_{F'|\tau} = \mathrm{E}_{X',y'\sim F'}\left(L\left(y', \hat{f}^{\tau}(X')\right)|\tau\right),$$

where here the expectation averages over draws from F'.

However, model validation captures differences due to *both* model fit issues (from estimating F by $\hat{f}^\tau$) and true differences between the DGMs (between F and F'). Instead of relying solely on validation measures, we propose using cross validation to properly calibrate the measurements from validation. In this way, we can separate the differences due to model fit issues and random variation (as measured by cross validation) from any true differences between the data generating mechanisms.

3.2. General framework

Letting $\kappa : (1, \dots, N) \to (1, \dots, K)$ be an indexing function specifying data splits, the estimate of predictive error given by K-fold cross validation is

$$\mathrm{CV}_\kappa(\hat{f}^\tau) = \frac{1}{N} \sum_{i=1}^{N} L\left(y_i, \hat{f}^\tau_{-\kappa(i)}\left(x_i\right)\right),$$

where κ is typically specified such that the number of observations in each of the K partitions is roughly equal. Cross validation estimates the *expected* test error, $\mathrm{E}_\tau[\mathrm{Err}_{F|\tau}]$.

Our prediction scores are designed to compare differences between validation and cross validation. In order to make this comparison meaningful, we need to consider a version of validation that estimates the conditional test error, $\mathrm{Err}_{F'|\tau}$, averaged over new potential draws of τ from F: $\mathrm{E}_\tau[\mathrm{Err}_{F'|\tau}]$. This can be achieved by redefining the typical validation loss statistics as follows:

$$\mathrm{Val}_\kappa(\hat{f}^\tau) = \frac{1}{N'} \sum_{i=1}^{N'} L\left(y'_i, \hat{f}^\tau_{-\kappa(i)}\left(x'_i\right)\right)$$

where the $\hat{f}^\tau_{-\kappa(i)}$'s are the *same* predictive models from cross validation (i.e., the models are trained on the same *subsets* of τ). This differs from the traditional implementation of validation in which the predictive model would be trained using all entries in τ. Intuitively, if this were the case, we would naturally expect better predictive performance in the validation routine, since the predictive model has the benefit of being trained on more data. Keeping the predictive models as comparable as possible (across the cross validation and validation routines) by training them on the same subsets of the data enables better detection of true differences between the underlying DGMs.

Differences between our redefined validation and cross validation estimate differences between the conditional test errors based on F and F', averaged over new training datasets τ. In other words,

$$\mathrm{Val}_\kappa(\hat{f}^\tau) - \mathrm{CV}_\kappa(\hat{f}^\tau) \text{ estimates } \mathrm{E}_\tau\left[\mathrm{Err}_{F'|\tau}\right] - \mathrm{E}_\tau\left[\mathrm{Err}_{F|\tau}\right]$$

and the estimand can be expanded as

$$\mathrm{E}_\tau\left(\int_{\mathcal{X},\mathcal{Y}} L\left(y, \hat{f}^\tau(x)\right) d\left(F(x,y|\tau) - F'(x,y|\tau)\right)\right),$$

which is the prediction error or loss averaged over differences between the DGMs and averaged over training sets τ. This expansion elucidates a natural connection; the differential resembles the form of the Kolmogorov-Smirnov statistic, a popular method that uses differences in empirical cumulative distribution functions (i.e., estimates for F and F') to measure discrepancies across univariate distributions.

In Algorithm 1 and Figure 2, we generalize the prediction scores so that comparisons between the cross validation and validation summary measures need not be computed only as the difference of means, as above. That is, our prediction scores are defined as

$$\Delta_{pred} = h\left(l, l'\right),$$

Algorithm 1 Prediction Scoring

Let $\tau = ((x_1, y_1), \dots (x_N, y_N))$ be a dataset with N observations drawn from the DGM F which is the joint population distribution for X and y; and analogously, τ' consists of N' observations and is drawn from F'. Let $\kappa : (1, \dots, N) \to (1, \dots, K)$ be an indexing function specifying data splits for τ. Let L be an appropriate loss function and let h be the prediction scoring function.

```
procedure Fit Predictive Models
    for $k = 1, \dots K$ do
        construct $\tau_{-k} = ((x_i, y_i)$ s.t. $\kappa(i) \neq k)$
        compute $\hat{f}^{\tau}_{-k}$, the predictive model trained on $\tau_{-k}$
procedure Cross Validation
    for $i = 1, \dots N$ do
        $l_i = L\left(y_i, \hat{f}^{\tau}_{-\kappa(i)}(x_i)\right)$
procedure Validation
    for $i = 1, \dots N'$ do
        $l'_i = L\left(y'_i, \hat{f}^{\tau}_{-\kappa(i)}(x'_i)\right)$
procedure Prediction Scoring
    $\Delta_{pred} = h\left(l, l'\right)$
```

where l and l' are the vectors of loss statistics such that $l_i = L\left(y_i, \hat{f}^{\tau}(x_i)\right)$ for $i = 1, \dots, N$ and $l'_i = L\left(y'_i, \hat{f}^{\tau}(x'_i)\right)$ for $i = 1, \dots, N'$; and h is a function that compares the distributions of loss statistics. Although in our definition above the prediction score is a *function* that compares the loss statistics, in practice, diagnostic plots that represent the differences between these distributions may be more useful, as demonstrated in later examples.

3.3. Additional considerations

Forecast distributions and non-Bayesian models. In order to appropriately account for uncertainty, we recommend making predictions in the form of vectors of possible outcomes, also called simulated forecast distributions. In the examples that follow we will adopt Bayesian predictive models which provide a natural way of computing vector-valued predictions; we can simply draw samples from the posterior predictive distribution for each observation. For non-Bayesian models, similar types of predictive distributions can be computed with bootstrapping or other resampling methods.

Choosing appropriate functionals. In practice, calculating prediction scores involves specifying a few important elements: the predictive model ($\hat{f}$), the loss function to compare predictions to realized data (L), and the subsampling method for the cross-validation and validation routines (κ). Additionally, to study the theoretical properties of the prediction scores, appropriate choices for d (the measure of the true difference between the data generating mechanisms) and h (the measure of the difference between the distributions of the loss statistics) must be made. These choices should be well motivated by the data types and modeling choices of the particular application. More specifically, the true DGM distance should be motivated by the form of the family of data generating mechanisms being considered, and the loss function should be motivated by the form of the chosen predictive model and model fitting software.

Because we encourage making predictions in the form of forecast distributions, the loss function needs to be capable of evaluating differences between the true observation, y_i, (a number) and the corresponding set of predictions (a vector). This still leaves questions as to the form of the loss in the face of different types of predictive models. For example, when using a linear regression model, predictions for y will be continuous and so quantiles may be a natural choice. However, for a logistic regression model, the predictions will be probabilities (between 0 and 1) while the observations are binary. Some variant of the area under the curve (AUC) statistic may be a better choice for L. Strictly proper scoring rules such as the logarithmic score or Brier score could be easily incorporated (Gneiting and Raftery, 2007).

Additionally, the choice of the prediction score, h, should be motivated by both of these considerations and the subsequent choices for d and L. Although this methodology would be simpler if d, L, and h were universally specified, it is important that they capture relevant features of the data generating mechanisms and are suitable to whatever modeling assumptions and model fitting software is chosen by the researcher. This sort of conditional specification is not unlike the choice of an appropriate link function for generalized linear models. This framework is nicely aligned with many popular measures of predictive accuracy. For example, choosing L as quadratic loss for a linear regression predictive model and choosing h appropriately will result in prediction scores that compare mean squared error across the cross validation and validation routines. In the examples discussed here we consider logistic regression predictive models and adopt the popular AUC statistic for L, examining histograms of these statistics in Section 5 and computing h as Kolmogorov-Smirnov statistics in Section 4. Deriving appropriate forms of d, L, and h for more dependent data, such as networks or time series, is an active area of future research. Ideally, the prediction scoring methodology, including these choices for d, L, and h, should be fully preregistered prior to any data collection. This does not preclude us from using the prediction scores *in an exploratory* fashion to discover interesting data features.

As demonstrated in Section 4, h can be specified as a test statistic (or p-value) from a nonparametric test of the hypothesis that the validation and cross-validation loss statistics come from the same distribution (e.g., the Kolmogorov-Smirnov test, the Mann-Whitney U test, or DeLong's test). However, some care should be taken when interpretting the results of these tests for real data. First, such tests generally assume independent, identically distributed samples whereas the groups of loss statistics under comparison are likely correlated; recall that the loss statistics are model-based predictions, and the validation and cross-validation routines use models trained on the same subsets of τ. Second, *any* difference between the cross-validation and validation loss statistics (i.e., between the preregistered hypothesis and the resulting experimental data) may be of substantive interest. For these reasons, we highly encourage using visual checks and diagnostic plots when evaluating differences between the loss statistics, as demonstrated in the following sections.

3.4. The predictive model as a lens

As we will emphasize in the following examples, the proposed prediction scores reveal differences between the DGMs along the dimension of the model used to make predictions. Consider evaluating differences between the same pair of DGMs in the face of two competing models. In the case where these predictive models are orthogonal in some sense (i.e., capture distinct features of the DGMs), we can imagine that each model should produce a set of prediction scores that capture differences between the DGMs only according to the features of the DGMs that each model is equipped to detect. For example, consider the illustrative diagram given in Figure 3. Here, we imagine two DGMs, F and F', which reside in a large, complex, multidimensional DGM space. The true distance between these DGMs, Δ_{DGM}, is typically unobservable, but we can calculate prediction scores relative to a model, which measure the distance between the distributions of cross validation and validation loss statistics. For example, using model $\hat{f}_1$, we can learn about the DGMs, F and F', by projecting them into a lower-dimensional space, the prediction space for model $\hat{f}_1$ (represented by the low-dimensional green rectangle in Figure 3). In this lower-dimensional space, we can measure the distance between the predictive accuracy of model $\hat{f}_1$ for data corresponding to F (this is represented in the empirical distribution for q_1 and is obtained via cross validation) and for data corresponding to F' (represented

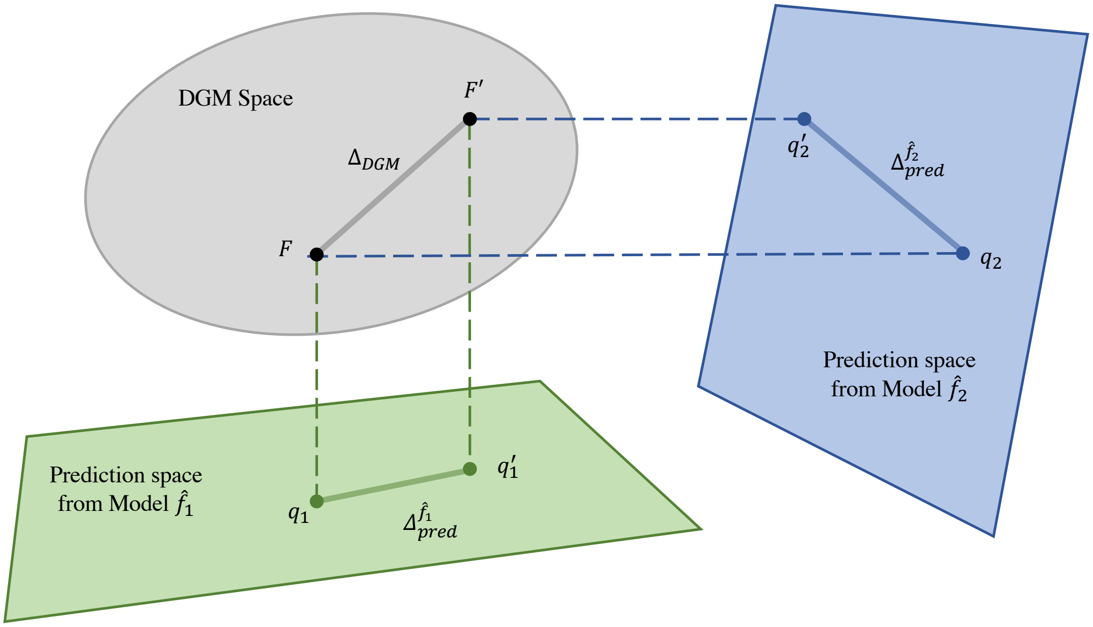
DGM Space
F'
Δ_{DGM}
F
q_2'
$\Delta_{pred}^{\hat{f}_2}$
q_2
Prediction space from Model $\hat{f}_2$
Prediction space from Model $\hat{f}_1$
q_1'
q_1
$\Delta_{pred}^{\hat{f}_1}$

Fig 3. *A geometric illustration of how prediction scores measure differences between DGMs along the dimension of the model used to make predictions.*

by q_1', obtained via validation). This prediction scoring distance, $\Delta_{pred}^{\hat{f}_1}$ depends on the model used to make predictions, $\hat{f}_1$, and will correspond to differences between F and F' that the model is equipped to detect. Now if we imagine a competing model, $\hat{f}_2$ (its prediction space is represented by the low-dimensional blue rectangle in Figure 3), which is "orthogonal" to this model, the resulting prediction scoring distance for model $\hat{f}_2$ will reflect different types of differences between the DGMs (i.e., in Figure 3, the blue distance is not the same as the green distance). In other words, DGMs may look more or less similar in terms of prediction scoring distances, depending on the model used to make the predictions. As we will discuss in more detail in Section 4, we can leverage this dependence and consider suites of "orthogonal" models in order to discover interesting differences between DGMs.

4. Simulated experiments

While our prediction scoring approach is motivated by the unique setting of preregistration, at its core it is a method for detecting differences between two data sources and their underlying DGMs. To demonstrate that these prediction scores can pick up on meaningful differences between competing DGMs (here, across experimental settings), we have designed a simulation study that utilizes a simplified experimental design, modeling the outcome of interest with logistic regression. The details of the simulated experiments are motivated by experimental data from a large-scale human behavior study, the Next Generation Social Science (NGS2) program (more details are available in the Appendix; experimental data from this program, and the motivating setting of preregistered hypotheses, will be explicitly considered in Section 5). This simulation study also allows us to examine how prediction scores can leverage competing predictive models to identify different types of differences between DGMs and to get a sense of their asymptotic behavior.

4.1. Experimental setup: human behavior in the presence of bots

Our simulation study will compare multiple settings of the simple and well-studied *public goods* game, from economic game theory (Ledyard, 1995). In a public goods game, n players have the opportunity to contribute ("cooperate") or not ("defect") over a series of T sequential rounds to a set of pooled resources that will be (multiplied and) shared among all participants. Each player's goal is to collect as many tokens (money) as possible; in each round, players are faced with the decision to be selfish (and keep their tokens), or be cooperative (and donate money to the common pool). In our simulated experiments, we assume that each player's decisions are made public to all other participants. Economists hypothesize that players' decisions to contribute at each round depends on the players' own baseline tendency to contribute, their previous decisions, and can be affected by the outcomes and behaviors of other players from the previous rounds.

Inspired by the experimental plan of the NGS2 research teams from the University of Pennsylvania and University of California, we imagine that bot-like participants play alongside the simulated human participants. Theoretically, inserting bot participants within these experiments would allow researchers greater experimental control over the social and environmental landscape within the game (Suchow et al., 2017) while simultaneously enabling the study of human behaviors in larger groups (i.e., adding bot participants is easier than recruiting human subjects). In this sense, researchers can use bot behaviors to create interventions and trigger different behaviors.

True DGM. In our simulation study, we consider an array of $K = 5$ different DGMs, representing different levels of the percent of bot participants in the game, from $\pi = (0, 0.25, 0.50, 0.75, 1)$. We are most interested in understanding the ways in which participants' decisions to cooperate are influenced by the presence of bots; thus, prediction scores will compare predictions for participant contribution across experimental settings where the percentage of bots differs.

For each experimental setting (i.e., each element of π), we imagine recruiting J cohorts of individuals to participate; let n_{jk} be the number of individuals competing in the jth cohort of the kth setting. Let y_{ijkt} be the decision to cooperate ($y_{ijkt} = 1$) or defect ($y_{ijkt} = 0$) for the ith individual in the jth cohort of the kth experimental setting during round t, where $i = 1, \dots n_{jk}, j = 1, \dots J, k = 1, \dots K$, and $t = 1, \dots T$. Additionally, let z_{ijk} be an indicator of whether the ith participant in the jth cohort of the kth round is a human participant ($z_{ijk} = 1$) or a bot (z_{ijk}). We will assume that $y_{ijkt}|z_{ijk}$ are independently distributed Bernoulli random variables, for all i, j, k, and t. Then, the true underlying data generating mechanism for the simulated data in our hypothetical experiments is given by the following:

$$\begin{aligned} & & z_{ijk} &\overset{iid}{\sim} \text{Bernoulli}(\pi_k) \\ \text{Model 0:} & & \text{logit}^{-1}\left(P(y_{ijkt} = 1|z_{ijk} = 1)\right) &= \beta_0 + \beta_1 t + \beta_2 y_{ijk,t-1} + \beta_3 \bar{y}_{\cdot jk,t-1} \\ & & \text{logit}^{-1}\left(P(y_{ijkt} = 1|z_{ijk} = 0)\right) &= \beta_0' + \beta_2' y_{ijk,t-1}, \end{aligned}$$

where π_k is the proportion of bots in the kth round, β_0 and β_0' are baseline tendencies to cooperate, β_1 captures any trend across the rounds, β_2 and β_2' capture the tendency to switch between behaviors, and β_3 represents the influence of team members' decisions. For example, if all other individuals cooperated in the previous round ($\bar{y}_{\cdot jk,t-1}$ is close to one), then the probability that individual i also cooperates in the next round is high, for large positive β_3. For bot participants, β_3 is defined to be zero; the simplistic bots we consider here are not influenced by the behavior of other participants. To specify reasonable parameter values for our simulation, we fit this true model to the experimental data from both Rand, Arbesman and Christakis (2011) and Diego-Rosell (2017) (analyzed in Section 5), using data from games played under the fluid network update setting for bot behavior and the fixed network setting for human behavior (see Table 1).

Parameter	β_0	β_1	β_2	β_3
Predictor	1	t	$y_{ijk,t-1}$	$\bar{y}_{\cdot jk,t-1}$
Human behavior	-1.31	-0.10	1.97	1.25
Bot behavior	-0.78	-	2.68	-

Table 1

Parameter values for Model 0: True data generating mechanism.

To mimic subject recruitment, for each setting, k, we will set the number of cohorts $J = 10$ and the number of rounds $T = 15$, and draw $n_{kj} \sim \text{Binomial}(M, p)$, where J and T are chosen to mimic the experimental settings specified by Rand, Arbesman and Christakis (2011) and Diego-Rosell (2017), $M = 10000$ is the size of the pool of possible recruits, and $p = 0.0018$ is the participation rate; this corresponds to roughly 18 participants per cohort and an expected 2700 data points per setting.

Prediction scoring. To mimic our analysis of experimental data in Section 5, we will fit Bayesian logistic regression models (this also aligns well with the true underlying DGMs). As discussed in Section 3.2, the form of the predictive model can help to inform our choice of loss function. Natural loss functions for these models include ROC or precision-recall curves.[2], and the corresponding area under the curve statistics. We will consider visual comparisons of these curves as well as differences in the distributions of the AUC statistics. Data subsets are created as random subsamples containing roughly 500 observations each (this translates to $K \approx 5$ subsamples per setting, since the expected sample size is 2700); since we are not investigating cohort- or individual-level effects of any kind, the data are partitioned completely randomly across all observations but is resampled in order to preserve consistent class proportions across all subsets. This resampling method is necessary to ensure that the precision-recall curves are comparable across datasets that vary by baseline cooperation rates (this is discussed in more detail in Section 5 and in Panel I of Figure 8).

Researcher models. We consider a suite of three potential researcher models:

$$
\begin{aligned}
&\text{Model 1:} \quad \text{logit}^{-1}\left[P(y_{ijkt} = 1)\right] &&= \gamma_0 + \gamma_1 t, \\
&\text{Model 2:} \quad \text{logit}^{-1}\left[P(y_{ijkt} = 1)\right] &&= \gamma_0' + \gamma_2 y_{ijk,t-1}, \\
&\text{Model 3:} \quad \text{logit}^{-1}\left[P(y_{ijkt} = 1)\right] &&= \gamma_0'' + \gamma_3 \bar{y}_{\cdot jk,t-1},
\end{aligned}
$$

where γ_0 is a baseline tendency to cooperate, γ_1 can capture some trends across the rounds, γ_2 represents the influence of the most recent decision, and γ_3 represents the influence of team members' decisions. In practice the true data generating mechanism is unknown to the researcher. However, the researcher typically has hypotheses about features of the DGMs that might differ across experimental settings and these features are incorporated in models as above. For example, if all participants are bots, than Model 2 should perform fairly well. However, whenever humans participate, Model 2 will fail to represent the full spectrum of observed behaviors well. The models specified here are intentionally non-nested. Since prediction scores are inherently model-based (i.e., they depend on the model used to make the predictions), recall from Figure 3 that we can interpret them as a distance between DGMs *along the dimension of a particular model.* In this sense, when trying to uncover features of the DGM that may differ across settings, models that can measure distinct features of the DGM should be prioritized. In some sense, we can think of the desired set of researcher models

[2] Generally, the precision-recall curve is preferred over the ROC curve when data are imbalanced, for example when there are many more 0's than 1's (see Davis and Goadrich, 2006, for more discussion) In our simulated data, even across each setting (i.e., where we compare data with π_{k_1} bot participants to data with π_{k_2} bot participants) the aggregate baseline cooperation rate varies from 0.44 (when both π_{k_1} and π_{k_2} are close to 0) to 0.66 (when both π_{k_1} and π_{k_2} are close to 1; these differences in the baseline rates are also apparent in the upper left panel of Figure 5).

as being orthogonal to each other[3] so as to maximize the possibility of discovering true differences between the DGMs.

4.2. Results

Shortcomings of predictive accuracy. First consider the more traditional approach of performing validation alone. In this case, the posterior predictive distribution is conditioned on the *full set of data* from the first experiment or dataset (predictors, x, and responses, y) but provides a prediction for the responses, y', from the second experiment, corresponding to the predictors in that second experiment, x'. This procedure is often used to compare competing models, such as those considered in our suite of researcher models here (see the far right panels in Figure 4). Whether the underlying DGMs differ or not (in the top row, both DGMs have $\pi_1 = \pi_2 = 0$ bots; the bottom row compares data from DGMs with π_1=0 and $\pi_2 = 0.50$), we observe that Model 3 has the best predictive accuracy. However, using these curves alone, it is impossible to say much about any underlying differences between the DGMs being compared. First, these curves fail to account for sampling variability. If the sample of participants in either experiment differed slightly, we would expect to see the curves in these figures move around a bit, but just how much they would move (i.e., how much sampling variability for this particular population or experiment impacts model fit and predictive ability) can not be estimated or accounted for in the validation-only procedure. In this sense, prediction scoring goes above and beyond traditional predictive accuracy measures; using subsets of the data in both the cross validation and validation routines helps to appropriately account for the effects of sampling variability. Secondly, these curves do not allow us to separate the effects of (poor) model fit from any true differences between the DGMs. Only by comparing cross validation curves to validation curves are we able to observe these differences. Both cross validation and validation curves are based on predictions made from the same model, so that any observed differences should solely reflect true differences between the DGMs.

Detecting DGM differences. First, consider the case where the DGMs are in fact identical across settings; see the top panel of Figure 4. As we would expect, there is little discernible difference between the cross validation and validation curves, regardless of the model used to make predictions. If instead we consider the case where there is a difference between the DGMs, such as $\pi_1 = 0$ and $\pi_2 = 0.50$ as in the bottom panel of Figure 4, we can see some evidence of a difference between the DGMs as we would expect. In short, the prediction scores are successful in detecting differences between the underlying DGMs.

Leveraging competing models. Recall from Figure 3 that the prediction scores are dependent on the predictive model and measure differences between DGMs along the dimension of the model used to make predictions. For Model 1, there is clear separation of the cross validation and validation curves indicating that there is a difference between these DGMs. For Models 2 and 3, this difference is less clear, but there does appear to be an ordering of the curves which is some indication of a difference across the two DGMs. Model 1 depends only on the round number. In fact, if we look at the raw data simulated for these experiments we see strong differences over time across these two settings. Thus it is not surprising that the prediction scores which come from a model that depends on time are particularly helpful in differentiating the two DGMs. In other words, the prediction scores reveal that the behavior of participants when there are 0% bots as compared to 50% bots differs most strongly with respect to the number of rounds; there is not a strong difference in regards to the participants' previous decisions or the average previous decision.

[3] Here we mean that the models should be non-nested, but we use the term "orthogonal" to better relate to the geometric description of prediction scores provided earlier—that they measure distance along the dimension of a particular model.

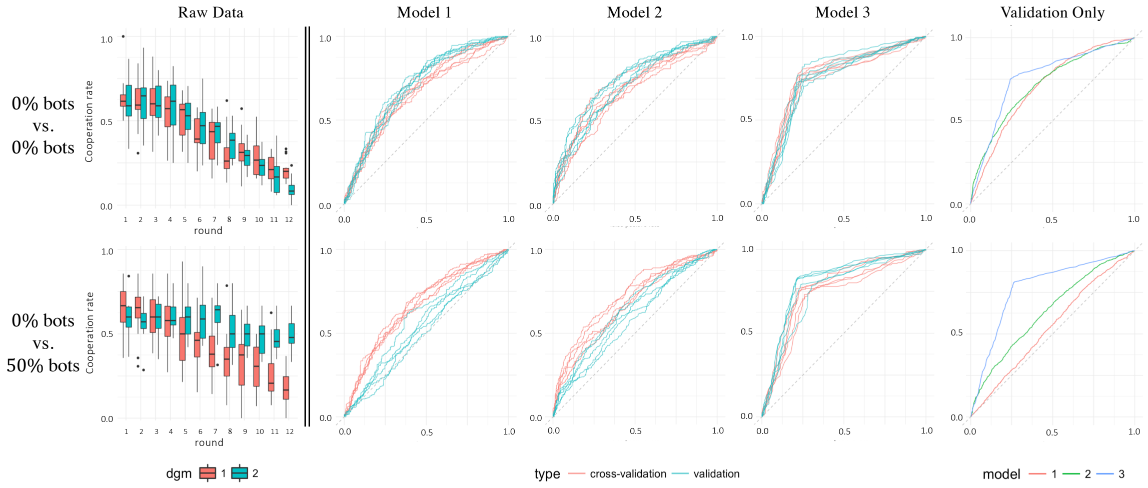

Fig 4. *ROC curves (true positive rate, y-axis, vs. false positive rate, x-axis) from prediction scoring to compare experimental settings. Across the top row, experimental settings are identical (i.e., the underlying DGMs are identical and so $F = F'$) with all human participants ($\pi = 0$) while in the bottom row, the experimental settings differ (i.e.,$F \neq F'$) and compares all human participants ($\pi_1 = 0$) to 50% bot participants ($\pi_2 = 0.50$). The plots on the far left contain boxplots of the cooperation rate across all cohorts and individuals by round, where the color represents the experimental settings. The plots on the far right are ROC curves for validation only, with each curve corresponding to a different researcher model. Remaining plots display the prediction scoring ROC curves for subsets of the data from the cross validation (red) and validation (blue) routines, with each plot corresponding to a different researcher model.*

Finally, we examine the prediction scores across a range of experimental settings, making comparisons across $\pi = (0, 0.25, 0.50, 0.75, 1)$, in Figure 5. Just as in Figure 4, we see that Model 1 is the most sensitive to differences across the experimental settings. Further, as we might expect, as the distance between the experiments increases (in terms of $|\pi_1 - \pi_2|$), so too does the separation between the cross validation and validation ROC curves, especially for Model 1. In other words, when the model is aligned with true differences between the data generating mechanisms, the distance between the cross validation and validation statistics reflects the true distance between the DGMs.

Summary. This simulation study demonstrates that prediction scores go above and beyond traditional predictive accuracy measures, can be used to uncover features of data generating mechanisms that differ across experimental settings, and can leverage competing predictive models to uncover different types of differences between the DGMs. This is true even when the true data generating mechanism is unknown, as is the typical case in practice. Here, the prediction scores discovered that the impact of round number or time in the game is best aligned with true differences between bot and human behavior. This is reassuring since we can verify this effect by examining boxplots of the cooperation rate by round across each setting (e.g., the leftmost panel in Figure 4).

4.3. Estimating prediction score accuracy

In order to get a sense of how these prediction scores behave asymptotically, we repeat the above simulation study and examine the relationship between the true distance between DGMs and our prediction scoring estimates of that distance. This requires defining a true distance between the data generating mechanisms. Here, we simply use the difference between the percentage of bots, $|\pi_i - \pi_j|$. We compare this true distance to the prediction scoring estimates of distance, which we

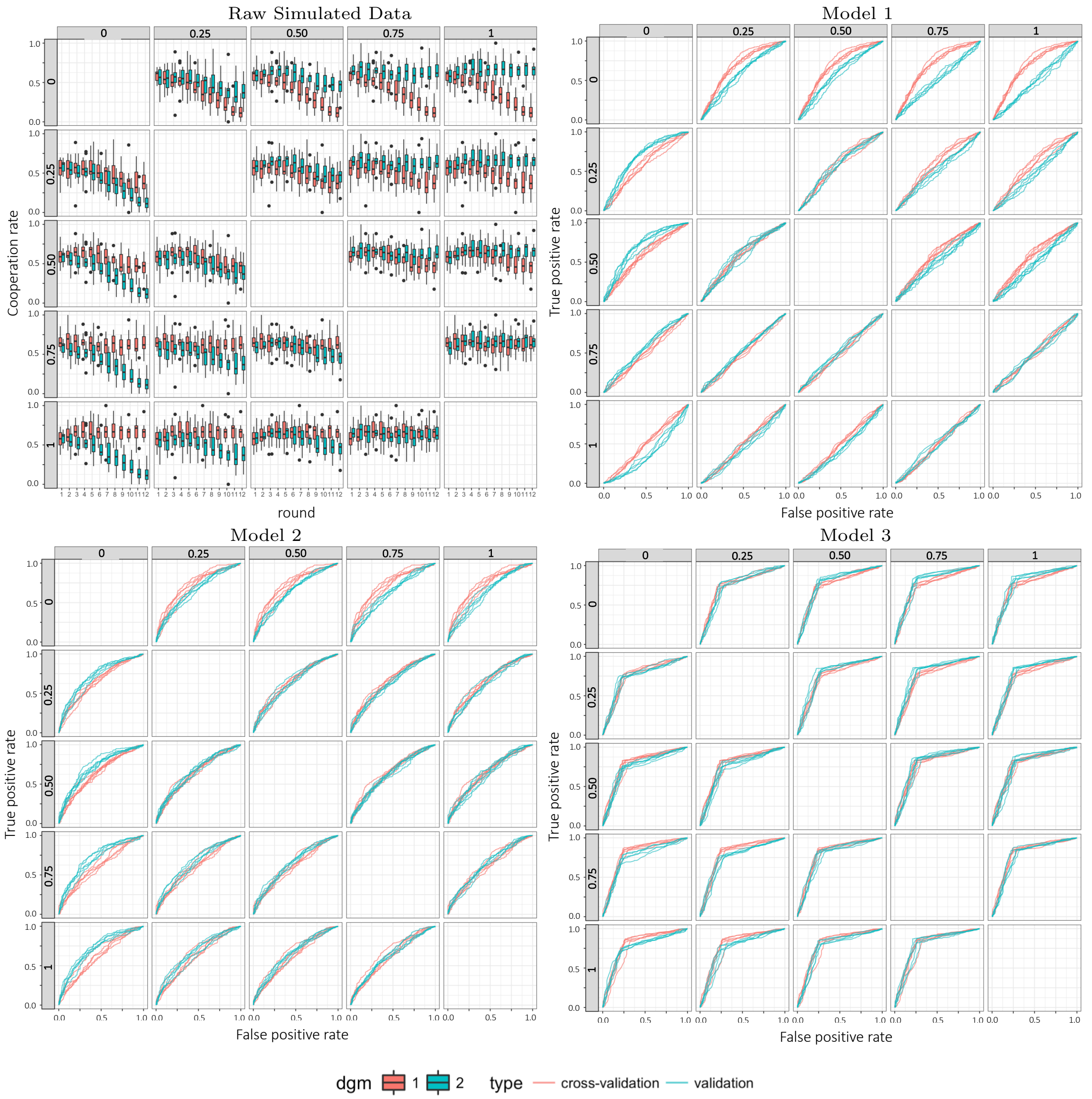

Fig 5. *ROC curves from prediction scoring to compare across experimental conditions for* $\pi \in (0, 0.25, 0.50, 0.75, 1)$. *The panel on the top left contains boxplots of the cooperation rate across all cohorts and individuals by round, where the color represents the experimental settings. Remaining panels display the prediction scoring ROC curves for subsets of the data from the cross validation (red) and validation (blue) routines, with each panel corresponding to a different researcher model. Within each panel, the columns correspond to experimental conditions (values of* π*) for the first DGM and the rows correspond to the second DGM.*

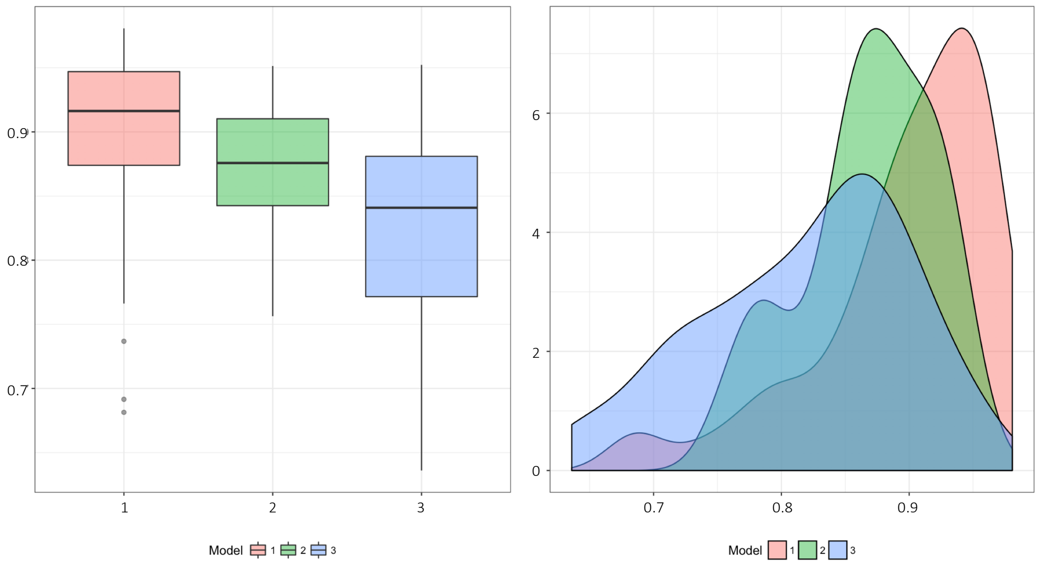

Fig 6. *Distance correlations for prediction scores, based on 100 repetitions of simulated experimental data. In each simulation, predictions are scored according to the predictive researcher models described earlier in Section 4 and by calculating Kolmogorov-Smirnov statistics that compare the empirical distributions of cross validation and validation AUC statistics. The left panel displays these results as boxplots; the right panel displays the same results as empirical density plots.*

calculate as Kolmogorov-Smirnov statistics (Kolmogorov, 1933) that compare the distributions of AUC statistics for the ROC curves across cross validation and validation. To evaluate whether or not the prediction scoring estimates are well-aligned with this measure of the true underlying distance, we calculate distance covariances (Székely, Rizzo and Bakirov, 2007). A distance covariance is a measure of dependence between two paired vectors that is capable of detecting both linear and nonlinear associations. If the vectors are independent, then the distance covariance is zero. We can treat each repetition of the above simulation study (where we compute prediction scores across all possible pairs of π) as a sample which gives rise to a vector of prediction scoring distance estimates. Then we examine the distribution of sampled distance covariances, as a function of the (researcher) model used to make predictions. After repeating this simulation 100 times, we plot the distance covariances in Figure 6. As expected, we see that on average Model 1 out-performs Model 2 which out-performs Model 3, in terms of how correlated the prediction scoring estimates of distances between the DGMs are with the true distance, as measured by the difference in the percentage of bot participants. This indicates that, on average, the prediction scores can successfully detect important differences between the DGMs.

5. Preregistered hypotheses in human behavior experiments

Recall from Section 2 that one important motivating example for prediction scoring methodology is in the case of evaluating preregistered hypotheses. In this section, we briefly review the idea behind preregistered analyses and traditional evaluation approaches. We evaluate a preregistered hypothesis against realized experimental data from Cycle 1 of the NGS2 program (see the Appendix for more details) from the research team led by scientists at Gallup (Diego-Rosell, 2017), demonstrating how prediction scores provide important advantages over traditional NHST procedures and that prediction scores can identify important differences between pilot and experimental data.

5.1. Preregistration

In a preregistered design, researchers prepare a detailed plan for all data collection, coding, and statistical analysis, along with the hypotheses and corresponding predictions regarding the study's results. This plan is made publicly available ("registered") in some way before ("pre") any data collection or analysis, so that the researchers are held accountable to their preregistered plan,[4] and the "garden of forking paths" can be safely avoided (Gelman and Loken, 2014). Preregistration ensures that in such settings where a p-value is useful, it can be interpreted correctly. Many journals, across many disciplines, now encourage preregistered studies, in the form of registered reports (e.g., the neuroscience journal, *Cortex*), and any study's preregistration materials can easily be made publicly available on sites like the Open Science Framework.[5]

As best we can tell, current practice for prediction scoring generally consists of making predictions in the form of directional hypotheses (in some cases, predictions for the relative effect size are also included) for parameters in a model that captures our beliefs about the true underlying DGM.[6] These predictions are then typically assessed by fitting the model to the observed experimental data, performing the corresponding hypothesis test and checking for a significant effect.

5.2. Example: Experimental human behavior data

Experimental setting. In this study, the Gallup team was interested in understanding the role of social networks in the public goods game (Ledyard, 1995). In this version, participants' contributions are split only among neighbors in their (possibly evolving) social network. Experimenters randomly assigned participants to one of four conditions which determined the dynamics of the social network in the game: (1) static or fixed links, (2) random link updating, where the entire network is regenerated at each round, (3) strategic link updating, where a randomly selected actor of a randomly selected pair may change the link status of that pair. The strategic link updating condition was further split into two categories: (a) viscous, where 10% of the subject pairs were selected and (b) fluid, where 30% of the subject pairs are selected. We will be primarily interested in the impact of the fluid version of the strategic link updating condition, from here on called "rapidly updating networks." The Gallup team used a logistic regression model to examine individuals' decisions (cooperation or defection) under a variety of experimental conditions. The Gallup team's experiments were inspired by the experiments performed by Rand, Arbesman and Christakis (2011) and whose data can serve as a set of preregistration pilot data.

Traditional analysis. We consider the following hypothesis from the Gallup team's preregistration materials: rapidly updating networks should support cooperation (across rounds of the game) more than any other condition (see Hypothesis 1.4 of Nosek et al., 2018). The traditional approach to evaluating this hypothesis would be to specify a model which includes a parameter that compares the rapidly updating network condition to all other conditions and then to perform a null hypothesis statistical test. Let y_{it} represent the decision to cooperate ($y_{it} = 1$) or defect ($y_{it} = 0$) for participant

[4] This sort of preregistration does not preclude further exploratory analyses; the point of preregistration is not to restrict analyses but rather to provide more structure to analyses that are already planned. For example, after data collection, a researcher may notice a pattern or posit a new explanation that motivates additional analyses. Such additional exploratory data analysis (beyond preregistered plans) are generally desirable as they can lead to new discoveries or hypotheses and even inspire additional confirmatory research.

[5] The preregistration materials corresponding to the study data used in the human behavior example discussed in Section 5 are hosted on this site.

[6] Other approaches include using Bayes factors or the small-telescopes approach (Simonsohn, 2015), though these methods seem far less popular than traditional NHSTs.

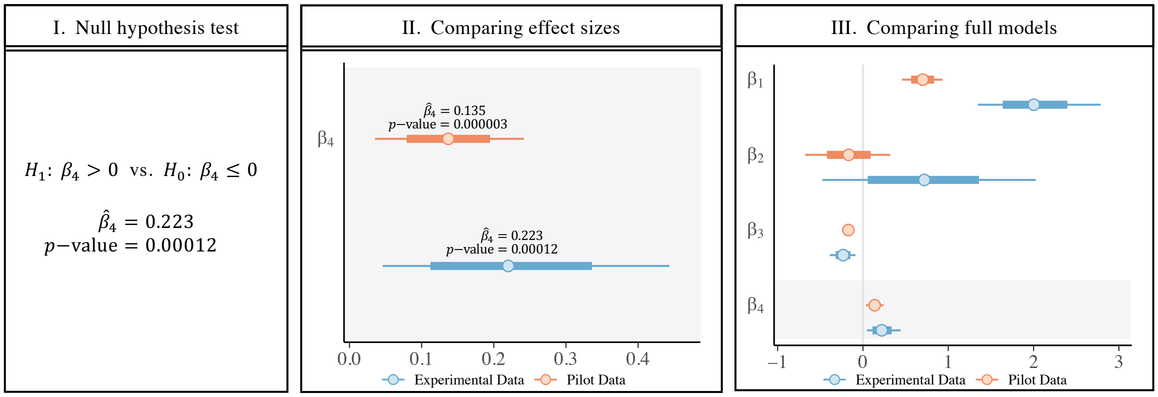

FIG 7. *Various approaches for evaluating a preregistered hypothesis in the face of pilot and experimental data.*

i in round t, which we can model as

$$y_{it} \overset{ind}{\sim} \text{Bernoulli}(p_{it})$$
$$\text{logit}^{-1}(p_{it}) = \beta_1 + \beta_2 t + \beta_3 X_i + \beta_4 X_i t,$$

where X_i represents inclusion in the rapid updating network condition for participant i. Then we are simply interested in testing:

$$H_1 : \beta_4 > 0 \qquad \text{vs.} \qquad H_0 : \beta_4 \leq 0.$$

In this case, the estimated effect size is about 0.22, with a p-value of 0.0001, and so one would traditionally conclude that the rapidly updating network conditions statistically significantly increase the likelihood of cooperation; see panel I of Figure 7. For this analysis and all those to follow, we match the modeling strategy proposed in Nosek et al. (2018) but will perform the analyses in a Bayesian setting.

However, this type of analysis does not incorporate the valuable information we have from the pilot data. Perhaps we could compare the results of this hypothesis test to its analogue for the pilot data; see panel II of Figure 7. In this case, the estimated effect size is about 0.14, with a p-value less than 10^{-5}. What does this allow us to say about how the pilot data compares to our experiment? In our pilot data there is a significant increase in the likelihood of cooperative behaviors under the rapid updating network condition; we find the same effect in our experiment, with a larger effect size but with slightly less evidence that this network condition makes a significant impact.

But this analysis ignores other trends in the data that might differ across the two data sources. Regression null hypothesis tests, like the ones above, are conditioned on all other predictors in the model. And if we compare these other effects across the two data sources (see panel III of Figure 7), we see a difference in the estimates for the baseline rapid updating condition. In our experiment, the baseline effect is positive and large and is significantly different than zero (though the p-value is on the larger side) whereas in the pilot data, the baseline effect is negative although not significant. While participants seemed to be much more cooperative overall in our experimental data, the way the game progresses also seems to affect the decision to cooperate across these two settings. From this analysis, it is unclear how these other differences between the experimental and pilot data might affect our hypothesis about the fluid network condition. Other summaries of model fit are generally unhelpful here as well; take for example, AIC, which can compare non-nested models (like these) but the interpretation of comparisons across different response data (i.e., different observed realizations of the outcome variable, y) is unclear.

So, how does the pilot data differ from our experiment? And how can we summarize these differences in a more holistic way that accounts for all trends related to our hypothesis about the rapid network condition? As in Section 4, we can consider traditional predictive accuracy metrics in the context of validation. As in Section 4, we will consider ROC and precision-recall curves,[7] and the corresponding area under the curve statistics. In this case, ROC seems to indicate that our model is little better than random guessing; see the yellow lines in panel I of Figure 8. At first the precision-recall curve appears to provide better news, but it is sensitive to differences in the proportions of 1's in the data. We also noticed this difference between the pilot and experimental data in our comparison of the full logistic regression models in panel III of Figure 7. So, to fairly assess the predictive accuracy of a model across two datasets, we need to ensure that the baseline rates are comparable. To accomplish this, we resample the experimental data so that the baseline rates across the datasets are matched (equivalent results can be obtained via reweighting). After this adjustment, the precision-recall curve seems to agree with the ROC in indicating that our model trained on the pilot data does not do a great job of predicting the experimental data; see the blue lines in panel I of Figure 8. Is that because our model does not represent the experimental data well (i.e., returning to Figure 2, the pilot and experimental data are clearly different and Δ_{DGM} is large)? Or could it be that our model doesn't represent either the pilot data or the experimental data well (i.e., our model doesn't represent the DGM family well, and so can't tell us much about Δ_{DGM}, the distance between the pilot and experimental data)? To answer these questions, we need to be able to assess how well our observed pilot and experimental data represent the underlying DGMs; we need to represent the variability across datasets generated from the same DGM. But these curves and any resulting analyses are conditioned on the particular observed (pilot) data. We have no way of understanding the inherent variability in these types of summaries. In fact, this is the case for all of the traditional analyses investigated thus far; null hypothesis tests, effect sizes and predictive accuracy measures are all conditioned on the particular set(s) of observed (training and testing) data.

Prediction scoring details. Recall from Figure 2, prediction scores require the following practical considerations: the predictive model, the loss statistic to compare predictions to realized data, and the subsampling method for the cross validation and validation routines. As mentioned above, we fit Bayesian logistic regression models from the Gallup team's preregistration materials and we calculate ROC and precision-recall curves along with their corresponding AUC statistics; this means that we are considering two different ways of scoring the predictions (in practice, there may be many appropriate loss statistics). Data subsets are created as random subsamples containing roughly 50 and 25 observations in training and testing sets, respectively. As discussed above, experimental data are resampled so that baseline rates across all datasets are comparable, which ensures that precision-recall curves can be accurately compared.

Results. We are evaluating the hypothesis that rapidly updating networks support cooperation more than any other condition. As in Section 4, we begin with visual comparisons of the ROC and precision-recall curves in panel III of Figure 8. In general, we see little separation between the cross validation and validation curves, indicating that the underlying DGMs are likely to be similar; this matches the conclusion from our NHSTs discussed above.

However, the prediction scores additionally reveal that the researchers' logistic regression model surprisingly is a slightly better fit to the experimental data (particularly in terms of ROC). This is somewhat surprising, since in most cases we expect a model to do a good job of predicting the data

[7] The precision-recall curve is preferred over the ROC curve when the data are imbalanced, typically when there are many more 0's than 1's (see Davis and Goadrich, 2006, for more discussion). This is not the case here, since there are $n_1 = 3876$ observations in the pilot data and the average decision to cooperate is $p_1 = 0.53$. Compare this to the experimental data, with $n_2 = 1192$ and $p_2 = 0.86$.

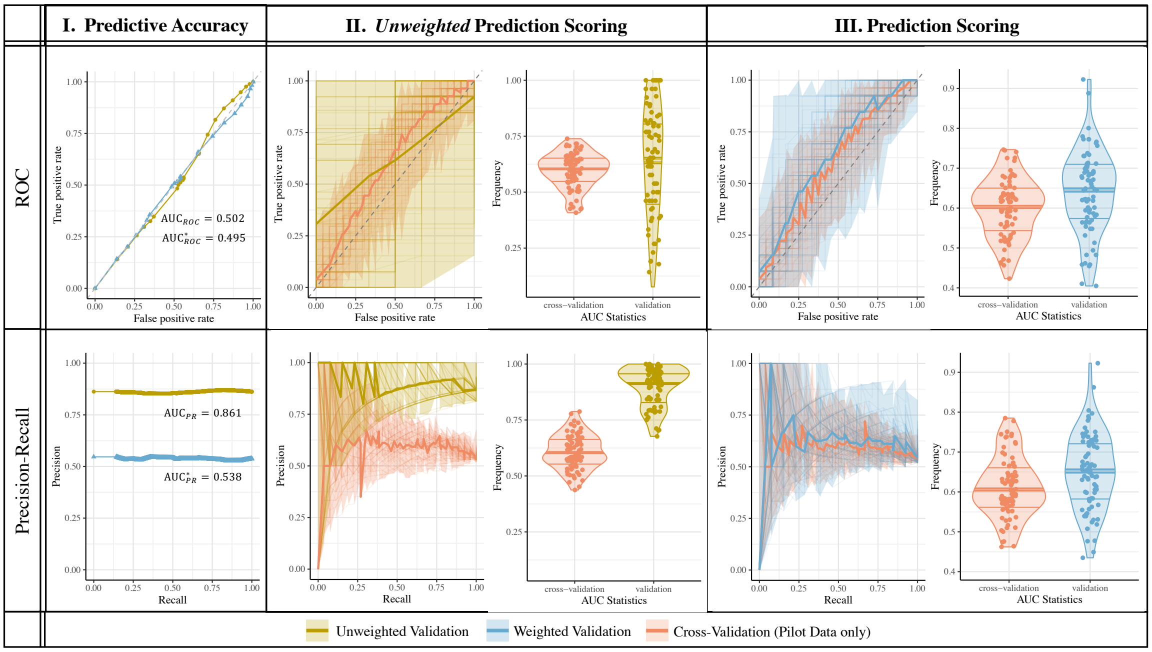

Fig 8. *Prediction scores for Gallup's Cycle 1 Hypothesis 1.4 that rapidly updating networks support cooperation, relative to all other conditions.*

it was trained with, which would result in cross validation curves that look better then validation curves. This indicates that there is less variability in individuals' behavior in the experimental data than in the preregistration data. In a sense, subjects in the Gallup experiment are acting in more predictable ways than the subjects in the pilot data. This conclusion is reached through the lens of the predictive model—here, logistic regression—and indicates that the experimental data are a better fit for this modeling framework (as opposed to the pilot data). That is, the prediction scores identify a way in which the underlying DGMs differ; using the notation in Figure 2, the underlying DGM for the experimental data, F', is more similar to a logistic regression model, $\hat{f}$, than the underlying DGM for the pilot data, F. And in fact, if we simply examine summary statistics of the in-game decisions themselves, we can see the same type of pattern. In Figure 9, we provide boxplots of individuals' average cooperation levels across rounds of the game, where each color corresponds to a different link-updating experimental condition. Comparing the preregistration data (top row) to the experimental data (bottom row), we see that the boxplots are drastically narrower, especially in the fluid network condition, indicating that there is less variability in participant behavior. This is not a difference between the two datasets that would have been picked up by the traditional prediction score, the p-value for the NHST associated with β_4.

The validation curves themselves vary more, across training sets. This is even more obvious if we compare the loss statistics, the AUC statistics for each of the curves in these figures. Again, we begin with a visual comparison of the distribution of AUC statistics from cross validation to those from validation; see panel III of Figure 8. Some of this could be due to the difference in sample sizes, particularly the difference in the sizes of the testing sets. In this analysis, we've chosen N/K, the size of the training sets, to be 50 observations; thus, with $N = 3876$ in the pilot data and $N' = 1915$ for the resampled experimental data (we need to resample here so that the baseline rate is comparable across both datasets), this yields testing sets of roughly 50 and 25 observations for the cross validation and validation routines, respectively. The preferred loss statistic will often dictate a

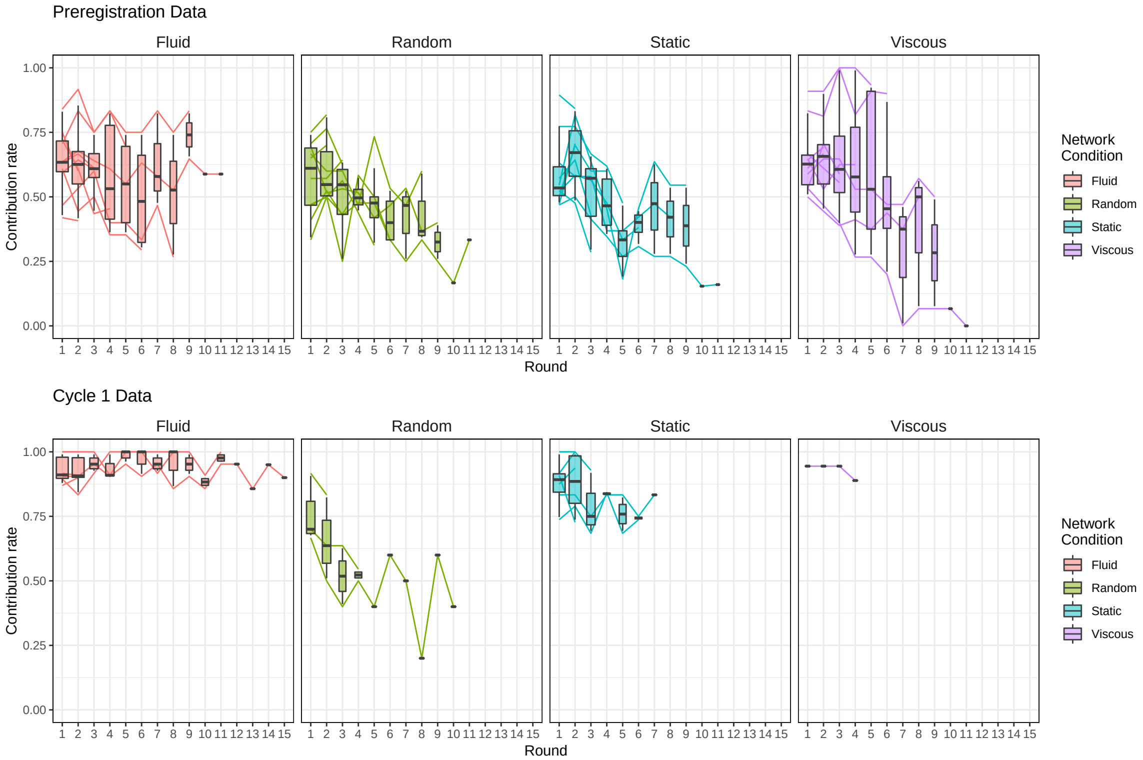

FIG 9. *Average cooperation levels across rounds of Gallup's Cycle 1 games.*

minimum size for these testing sets. So, here, it is possible that some of this increased variability in the validation curves is due to the smaller testing set size. To counter this effect, we could consider resampling the experimental data further, to match the size of the pilot data. But this runs the risk of going too far in the opposite direction, as resampled data tends to underestimate sampling variability.

In addition to the visual observations above, we could also consider performing some hypothesis tests of whether or not the cross-validation loss statistics and the validation loss statistics appear to come from the same distribution, indicating that, through the lens of the chosen predictive model, the underlying DGMs are indistinguishable. As discussed at the end of Section 3.3, the results of such tests should be interpreted cautiously, as the samples of loss statistics are likely correlated and *any* difference between the loss statistics could be of practical significance to substantive researchers. In this case, the Kolmogorov Smirnov test statistic for equality of distributions of the weighted cross-validation and validation loss statistics is $D = 0.325$, (p-value $= 0.000543$),

We have created similar summary measures for the unweighted version of the prediction scores; see panel II of Figure 8. As mentioned earlier, the precision-recall curves are much more sensitive to the difference in the baseline rates across the two datasets. This is not an obvious consequence of the precision-recall curves themselves, so care must be taken in fine-tuning the prediction scoring algorithm (i.e., choosing an appropriate loss statistic and possibly adjusting the resampling method) in order to obtain prediction scores that capture interesting features of the underlying DGMs, rather than more basic differences of the observed data (e.g., baseline cooperation rates). In this sense, it can be helpful to consider competing loss statistics, as shown here with AUC statistics for both ROC and precision-recall curves.

Overall this application serves as an illustration of how our prediction scoring method can be used to evaluate preregistered hypotheses and to enable interesting scientific insights. Here, while the NHST confirmed our preregistered hypothesis (the fluid network condition does support cooperation), prediction scoring nicely complemented this analysis by confirming that the DGMs appear to be similar and additionally highlighting differences in the underlying DGMs through the lens of our logistic regression model: participants in the experimental data acted more predictably than participants in the pilot data, particularly in the fluid network condition. In other words, the experimental data exhibited less variation in participants' decisions to cooperate over the course of the game (after accounting for differences across rounds).

6. Discussion

A natural version of the *prediction scoring* question might be phrased as follows: are these experiments or realizations products of the same DGM or are they distinct in some way? While this is certainly a natural question, it is ill-posed for most experimental research settings. In almost all cases, the DGMs do in fact vary across experiments or settings (e.g., from preregistration to observed data), even if only slightly. Instead, we have focused on answering the following question: How much do the data generating mechanisms differ across settings? To this end, our methodology provides prediction scores on a continuous scale; these scores can be viewed as estimates of the distance between our prior beliefs and reality. Thus, they provide a quantitative measure of how well preregistered predictions align with reality rather than relying on the simple binary detection of a significant effect.

This methodology utilizes cross validation and model-based predictions to solve a common problem in applied statistics research: the evaluation of differences between DGMs. In practice, the DGMs may represent related experiments, different settings or conditions within a single experiment, or preregistered hypotheses and realized observational data. We argued that comparing DGMs should move away from the simplistic binary question of whether or not the DGMs are equal and instead our prediction scores enable a quantifiable measurement of differences between DGMs. In an application to human behavior experiments, we demonstrated how the prediction scores can be used to evaluate preregistered hypotheses and for a set of simulated experimental data, we demonstrated how these scores can detect important differences between experimental settings. We also provided some intuition for the probabilistic behavior of these scores in an asymptotic regime.

Our application to the NGS2 project highlights the role our proposed prediction scores can play in light of the replication crisis. The majority of statistical concerns arising from the replication crisis fall into one of the three following topics: (1) selection bias (i.e., only research with small p-values is submitted and published), (2) insufficient power (i.e., p-values that correspond to small samples are not to be trusted), and (3) how statistical inference differs from scientific inference (Colling and Szűcs, 2018). Our prediction scoring methodology attempts to address this third issue but cannot fix issues arising from selection bias or insufficient power; these concerns still need to be carefully monitored. Much of the discussion around this issue has focused on reconciling the questions scientists want answers to with the questions that traditional p-values are equipped to provide. The proposed prediction scores are not p-value replacements; they are designed to answer yet another sort of scientific question: how do we quantify differences between DGMs? Or in the preregistration setting, how do we compare preregistered hypotheses to realized experimental data in a meaningful way? At its core, prediction scoring is a method for quantifying the distance between competing DGMs, one representing our prior/preregistered beliefs and one representing the realized experimental data.

In the examples in this paper, we focused on (visually) identifying differences between the distributions of loss statistics. In practice, visual inspection of plots always has a subjective component;

these plots also represent empirical samples of the full loss statistic distribution (across infinitely many subsampled testing/training sets). Even when the prediction scoring procedure itself is pre-registered, this interpretation—which cannot be fully pre-registered because it depends on realized experimental data—involves some amount of subjectivity or researcher degrees of freedom. However, as mentioned above, our proposed approach is designed not as a p-value replacement but rather as a tool to take advantage of the statistical problem presented by the preregistration setting: to provide researchers with a method to make these valuable comparisons between pre-experimental data and realized observed data in a way that goes beyond simple NHSTs.

As proposed, our prediction scores are not proper distances. In this sense, we don't expect them to be symmetric. It matters which dataset plays the role of τ (i.e., is used to create a baseline, via cross validation) and which plays the role of τ' (i.e., is used in validation). In the motivating example of the preregistration setting and as discussed in the real data example in Section 5, there is a natural directionality. However, in other settings this may not be the case. For more general cases, perhaps a symmetrized distance could be created by swapping the roles of τ and τ', and averaging the resulting sets of scores in some fashion.

AUC statistics were utilized in all of the case studies investigated here. It is worth noticing that the traditional AUC statistic is not defined for datasets consisting of only one observation. This means that leave-one-out cross validation is not feasible for this choice of loss statistic. In our applications, we accommodated for this by choosing $k < n$ for the k-fold cross validation routines in our prediction scoring method; this ensures that there are sufficient data points in the holdout sample. Alternatively, other approaches for calculating the ROC or precision-recall curves in small sample settings could be incorporated (e.g., Yousef, Wagner and Loew, 2005).

Both the application and simulated experiments involve rather simple DGMs and predictive models (i.e., both involve logistic regression models). This is by design, as we would like to be able to verify that the prediction scores are capable of detecting meaningful differences between the DGMs. However, we recognize that many applied problems hope to address more complicated DGMs and require more complex modeling strategies. This is an important avenue for future research; as discussed earlier, prediction scores are well-equipped to evaluate complex DGMs since they do not rely on the choice of a single parameter or summary statistic upon which to base the evaluation of differences between the DGMs. For example, in planned future experiments from the NGS2 program, researchers have preregistered hypotheses based on Gaussian process models for the DGMs under study. We hope to utilize prediction scoring in these settings, both to evaluate the preregistered hypotheses and to uncover interesting differences between experimental settings. For example, in some experiments, bots (computer agents whose behavior is algorithmically determined) will participate alongside human participants; we hope to use prediction scores to detect scientifically interesting differences in the observed (highly nuanced) human behavior patterns between human-only and human-and-bot experimental conditions.

Funding

This material is based on and supported through research sponsored by the Defense Advanced Research Projects Agency (DARPA) agreement number D17AC00001 and the Office of Naval Research. The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

The authors have no relevant financial or non-financial interests to disclose.

Data and Code

Experimental data analyzed in Section 5 is available on the Open Science Framework through associated GitHub links (Diego-Rosell, 2017). The simulated data and all code used in this paper are available in an additional public Github repository (Smith, 2020).

Appendix A: The NGS2 Program

The NGS2 program is a multi-phase methodologically-focused effort to develop a fundamental reimagining of the social science research cycle (Nosek et al., 2018). In each phase, distinct research teams conducted unique experimental social science studies regarding a shared research question. Prior to any data collection, each team was required to document all preregistration materials, including predictions for study outcomes (for more detailed descriptions of each team's planned and completed research, see the preregistration materials which have been made publicly available on the Open Science Framework; Nosek et al., 2018). The program also required that each team's preregistered hypotheses and resulting final analyses be evaluated by external non-team members, which included the authors of this paper. It is precisely in this context that our proposed prediction scoring methodology was developed.

In the first cycle of the program, research teams focused on explaining and predicting the emergence of collective identities. Collective identity refers to the way individuals perceive themselves in their environment with respect to the various groups they may belong to and how they subsequently take collective action or display collective behaviors related to this identity. In Section 5, we examine how well the preregistered hypotheses align with the realized experimental data from the research team led by scientists at Gallup (Diego-Rosell, 2017). The Gallup team's experiments were designed to mimic those of Rand, Arbesman and Christakis (2011); these experiments serve as pilot data and helped to formulate the team's preregistered hypotheses.

In the second cycle of the program, research teams designed experiments and analyses to study the emergence of group innovation in the face of competition. The design of the simulation study described in this section is inspired by the proposed Cycle 2 experiments of the research team led by scientists at the University of Pennsylvania (Suchow et al., 2017). These experiments examine human behavior in the face of computer generated participants.

Appendix B: Additional background

In this section, we provide relevant background information on the existing statistical methods which motivate our proposed prediction scoring framework, which is fully developed in Section 3. As outlined previously, our approach is inspired by existing tools for measuring predictive accuracy (Section B.1) and proposes a framework based on cross validation (Section B.2). We also offer a discussion of how our proposed approach relates to recent work in algorithm validation (Section B.3), which motivates strategies adopted in our proposed framework.

B.1. Scoring rules

Scoring rules measure the agreement between a probabilistic forecast (a predictive probability distribution over future quantities or events of interest, such as a posterior predictive density from a Bayesian analysis) and an observation (Gneiting and Katzfuss, 2014, provides a nice summary of recent research in this area). This literature provides a sound framework for comparing probabilistic forecasts or predictions (such as from preregistration materials) to observed data, where each competing forecast could correspond to different modeling choices or assumptions. The diagnostic tools and recommendations for scoring rules—e.g., checking for uniformity in histograms (or empirical CDFs, if the sample size is small) of the PIT (probability integral transform) values (this idea can be traced as far back as Rosenblatt, 1952; Pearson, 1933, and perhaps earlier)—are predictive and thus enable the comparison of non-nested, highly diverse models fit to common data. For example, Pers et al. (2009) use strictly proper scoring rules to select between competing machine learning models. However, since these tools were developed from the perspective of forecast selection (e.g., choosing the best forecast from among a group of competing forecasts), each set of resulting diagnostic measures is necessarily model-based in that any diagnostic plot or set of scoring rules depends on the model assumptions used to create the probabilistic forecast. This complicates the interpretation of the scores or diagnostics in regards to true underlying differences between the DGMs, since they can detect differences between the DGMs but are also designed to measure differences between the model and the DGM, which may be attributable to model fit issues. As mentioned earlier, our proposed prediction scoring approach uses cross validation to help normalize for model fit issues, and many of the proposed scoring rules could be incorporated in our proposed method.

B.2. Cross validation

Cross validation, particularly for Bayesian analyses, has been an active research area in recent years. Summary statistics for comparing Bayesian models can be motivated by estimation of out-of-sample predictive accuracy (see Vehtari et al., 2012, for a thorough review, from a formal decision theoretic perspective), which is one of the goals of cross validation as well. Gelman, Hwang and Vehtari (2014) provide a review of some model comparison summary measures, including AIC, DIC, WAIC, in the context of Bayesian model comparison. As opposed to exact leave-one-out cross validation (LOOCV), each of the Bayesian model summary statistics utilize the full predictive density and perform an adjustment (e.g., importance sampling, or division by an appropriate variance) to remove the effect of over-fitting, since no data was actually held out. The authors conclude the paper by citing cross validation as their preferred method for model comparison, despite its high computational cost and requirement that data can be easily partitioned (i.e., partitioning is often not straight forward for dependent or hierarchical data). In this line of thought, Vehtari, Gelman and Gabry (2017) develop an approximate version of leave-one-out cross validation which implements Pareto smoothing of the importance sampling weights to improve robustness to weak priors or influential observations. Li et al. (2016) develop a version of cross validation that can be applied to models with latent variables, which relies on an integrated predictive density. In applications with competing probabilistic forecasts, Held, Schrödle and Rue (2010) compare software fitting algorithms using approximate cross validation and many of the diagnostic plots mentioned by Gneiting, Balabdaoui and Raftery (2007). Finally, Wang and Gelman (2014) and Millar (2018) address the problem of appropriate data partitioning and out-of-sample prediction error estimation for multilevel or hierarchical model selection using cross validation and predictive accuracy. Wang and Gelman (2014) highlight the fact that model selection can be largely based on the size and structure of the hierarchical data.

This line of research, and its proposed improvements and extensions of cross validation in various Bayesian settings, can certainly be incorporated in the prediction scoring methodology that we propose. Our proposed approach expands this literature, from the perspective of the preregistration setting; we formalize the use of cross validation to appropriately adjust agreement measures between preregistered predictions and realized observations. In other words, we recommend a unique combination of cross validation *and* external validation to provide meaningful prediction scores.

B.3. Algorithm checking

Although perhaps not obvious at first glance, recent proposals for checking algorithms of Bayesian model fitting software (Cook, Gelman and Rubin, 2006; Talts et al., 2020) can provide insights in the prediction scoring setting. These proposals recommend simulating fake data conditional on random draws from the prior distribution, running the model fitting software to obtain draws from the posterior distribution, and using a summary measure to diagnose the alignment between posterior samples and the random draws from the prior distribution. Based on the self-consistency property of the marginal posterior and the prior distribution, these draws should be indistinguishable from one another. To diagnose this alignment, Talts et al. (2020) suggest computing rank statistics, comparing the random draw form the prior distribution to the posterior distribution based on that particular draw. The authors suggest looking at histograms of these quantiles, demonstrating that if the software is working correctly, the quantiles should follow a discrete uniform distribution. In the prediction scoring setting, we can think of this software-checking methodology as a special case where the chosen modeling strategy matches the underlying DGM exactly (i.e., there are no model fit issues whatsoever). We will borrow ideas from this methodology, such as the use of empirical quantiles and rank statistics to compare DGMs (or distributions) through samples drawn from them.

References

Billheimer, D. (2019). Predictive inference and scientific reproducibility. *American Statistician* **73** 291–295.

Colling, L. J. and Szűcs, D. (2018). Statistical inference and the replication crisis. *Review of Philosophy and Psychology* 1–27.

Cook, S. R., Gelman, A. and Rubin, D. B. (2006). Validation of software for Bayesian models using posterior quantiles. *Journal of Computational and Graphical Statistics* **15** 675–692.

Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning* 233–240. ACM.

Diego-Rosell, P. (2017). Experiment 1. Open Science Framework. osf.io/6jvw9.

Geisser, S. (2017). *Predictive Inference.* Routledge.

Gelman, A. (2006). Multilevel (hierarchical) modeling: What it can and cannot do. *Technometrics* **48** 432–435.

Gelman, A. (2013). Preregistration of studies and mock reports. *Political Analysis* **21** 40–41.

Gelman, A., Hwang, J. and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing* **24** 997–1016.

Gelman, A. and Loken, E. (2014). The statistical crisis in science. *American Scientist* **102** 460 - 465.

Gneiting, T., Balabdaoui, F. and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B* **69** 243–268.

Gneiting, T. and Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and its Application* **1** 125–151.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102** 359–378.

Hastie, T., Tibshirani, R. and Friedman, J. (2017). *The Elements of Statistical Learning, second edition.* Springer.

Held, L., Schrödle, B. and Rue, H. (2010). Posterior and cross-validatory predictive checks: A comparison of MCMC and INLA. In *Statistical Modelling and Regression Structures* 91–110. Springer.

Humphreys, M., Sanchez de la Sierra, R. and Van der Windt, P. (2013). Fishing, commitment, and communication: A proposal for comprehensive nonbinding research registration. *Political Analysis* **21** 1–20.

Jeske, D. (2019). Statistical inference in the 21st century: A world beyond $p < 0.05$ [special issue]. *American Statistician* **73**.

Kolmogorov, A. N. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari* **4** 1-11.

Ledyard, J. O. (1995). Public goods: Some experimental results. In *Handbook of Experimental Economics* (J. Kagel and A. Roth, eds.) Princeton University Press.

Li, L., Qiu, S., Zhang, B. and Feng, C. X. (2016). Approximating cross-validatory predictive evaluation in Bayesian latent variable models with integrated IS and WAIC. *Statistics and Computing* **26** 881–897.

Millar, R. B. (2018). Conditional vs. marginal estimation of the predictive loss of hierarchical models using WAIC and cross-validation. *Statistics and Computing* **28** 375–385.

Nosek, B. A., Spitzer, M., Russell, A., Tully, E., Rajtmajer, S., Ahn, S.-H., Zheng, T., Foy, D., Kluch, S. P., Stewart, C. and et al. (2018). NGS2 DARPA Program. Open Science Framework. osf.io/4jbx4.

Nuzzo, R. (2014). Statistical errors: P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume. *Nature* **506** 150–153.

Pawel, S. and Held, L. (2020). The sceptical Bayes factor for the assessment of replication success. *arXiv:2009.01520.*

Pearson, K. (1933). On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. *Biometrika* **25** 379–410.

Pers, T. H., Albrechtsen, A., Holst, C., Sørensen, T. I. and Gerds, T. A. (2009). The validation and assessment of machine learning: a game of prediction from high-dimensional data. *PLoS One* **4** e6287.

Racine, J. (2000). Consistent cross-validatory model-selection for dependent data: *hv*-block cross-validation. *Journal of Econometrics* **99** 39–61.

Rand, D. G., Arbesman, S. and Christakis, N. A. (2011). Dynamic social networks promote cooperation in experiments with humans. *Proceedings of the National Academy of Sciences* **108** 19193–19198.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W. et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **40** 913–929.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *Annals of Mathematical Statistics* **23** 470–472.

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science* **26** 559–569.

Smith, A. L. (2020). PredictionScoring. Github repository. https://github.com/annalantzsmith/predictionscoring.

Suchow, J. W., Stewart, A. J., Morgan, T. J. H., Malkomes, G., Krafft, P., Lall, V., Mosleh, M., Arechar, A., Akcay, E., Morsky, B., Rand, D., Plotkin, J. B. and Griffiths, T. L. (2017). Innovation in adversarial collective-sensing game. Open Science Framework. osf.io/zpvd3.

Székely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Annals of Statistics* **35** 2769–2794.

Talts, S., Betancourt, M., Simpson, D., Vehtari, A. and Gelman, A.

(2020). Validating Bayesian inference algorithms with simulation-based calibration. http://www.stat.columbia.edu/ gelman/research/unpublished/sbc.pdf.

Tukey, J. W. (1972). Data analysis, computation and mathematics. *Quarterly of Applied Mathematics* **30** 51–65.

Vehtari, A., Gelman, A. and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* **27** 1413–1432.

Vehtari, A., Ojanen, J. et al. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys* **6** 142–228.

Wang, W. and Gelman, A. (2014). Difficulty of selecting among multilevel models using predictive accuracy. *Statistics and its Interface* **7** 1–88.

Wasserstein, R. L. and Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *American Statistician* **70** 129–133.

Yousef, W. A., Wagner, R. F. and Loew, M. H. (2005). Estimating the uncertainty in the estimated mean area under the ROC curve of a classifier. *Pattern Recognition Letters* **26** 2600–2610.

Ziliak, S. and McCloskey, D. N. (2008). *The Cult of Statistical Significance: How the Standard Error Costs us Jobs, Justice, and Lives.* University of Michigan Press.