	All Games		Lab Games	
	Accuracy	Completeness	Accuracy	Completeness
Guess at random	0.33	0%	0.33	0%
PDNE	0.56	34%	0.38	7%
Level-1(α)	0.68	52%	0.79	69%
	(0.02)		(0.02)	
Level-1(α) + PDNE	0.79	69%	0.82	73%
	(0.03)		(0.03)	
Ideal prediction	1	100%	1	100%

Table 7: The level-1(α) + PDNE hybrid model improves upon the performance of both component models.

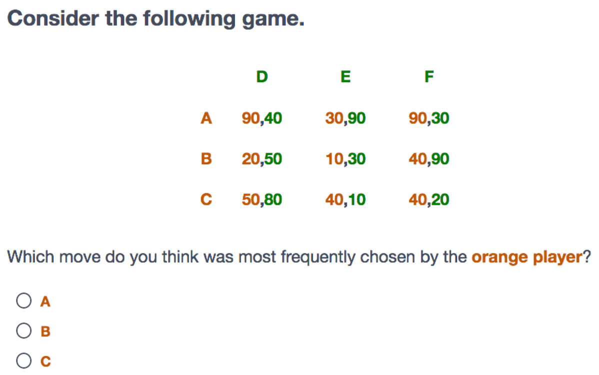
Our analysis above considers a specific hybrid model that combines two interpretable models. In principle, hybrid models can be built from a wide array of component models. For example, instead of combining two behavioral/economic models as we do here, we could combine a model such as level-1(α) with an algorithmic model, such as lasso or logistic regression. This kind of model would further blur the distinction between "behavioral" and "algorithmic" approaches. For more complex problem domains, such as predicting the distribution of play, we might consider hybrid models that combine two different structural models of play—for example, PCHM and a mixture-model of level-k types (as in Costa-Gomes, Crawford and Broseta (2001)). Yet another possibility is to combine a model based on the game matrix (as all of the approaches discussed so far are) with more "unconventional" models that use auxiliary data. We pursue this option below by using human forecasts as one component of the hybrid model.

7 Crowd-Sourced Forecasts

7.1 Human Predictions

We asked human subjects on Mechanical Turk to predict the most likely action in the laboratory games and algorithmically generated games.[47] We informed subjects that these games had been played by real people, and asked them to predict the action that was most likely to be chosen by the row player. On top of a base payment of \$0.25, subjects received an additional \$0.10 for every question they answered correctly. Figure 7 shows a typical question prompt presented to subjects, and the complete set of instructions can be found in Appendix F.

[47] Subjects were not screened based on level of exposure to game theory. The vast majority of answers suggest a lack of prior exposure to game theory, but some subjects did use terminology such as "dominance" in their post-survey responses (see Appendix G). The initial part of our experiment consisted of an introduction to matrix games, and we allowed subjects to proceed to the main experiment only after correctly reporting the payoffs for a fixed action profile in two example matrices (see Appendix F). All subjects eventually answered both comprehension questions correctly.



Consider the following game.

	D	E	F
A	90,40	30,90	90,30
B	20,50	10,30	40,90
C	50,80	40,10	40,20

Which move do you think was most frequently chosen by the **orange player**?

○ A
○ B
○ C

Figure 7: A typical question prompt presented to Mechanical Turk subjects in the single action treatment. The “orange player" is the row player.

Our experiment generated 40 crowd predictions for each game. We first consider the most direct use of these crowd predictions, which is to predict that the modal action is the most popular crowd prediction. We call this the *crowd forecast*. Table 8 shows that this simple crowd forecast performs remarkably well, improving upon the performance of the decision tree, PDNE, and level-1(α) for predicting play in the set of all games.

	Accuracy	Completeness
Guess at random	0.33	0%
PDNE	0.56	34%
Level-1(α)	0.68	52%
	(0.02)	
Decision Tree	0.70	55%
	(0.03)	
Crowd	0.77	66%
Ideal prediction	1	100%

Table 8: Crowd forecasts are predictive.

One potential explanation for the performance of the crowd forecast is that subjects predict the actions that they themselves would choose. This hypothesis would imply that each prediction is equivalent to an observation of play, so that with sufficiently many predictions, the distribution of crowd predictions would approximate the distribution of play arbitrarily well. We show below that this is not a complete explanation of the performance of the crowd forecasts.

7.2 Do People Predict Their Own Play?

Below we compare the distributions of play with the distributions of crowd predictions. Formally, we conduct chi-squared tests of the null hypothesis that our samples of game play and samples of crowd predictions are drawn from the same distribution. If the crowd predictions and game play were indeed drawn from the same distribution in every game, then the p-values for the chi-squared test would follow a uniform distribution. But we reject (under a Kolmogorov-Smirnov test) that the distribution of p-values is uniform with $p \approx 10^{-8}$ for the lab games, $p = 0.0027$ for the randomly-generated games, and $p \approx 10^{-15}$ for the algorithmically-generated games (see Figure 8 below).[48]

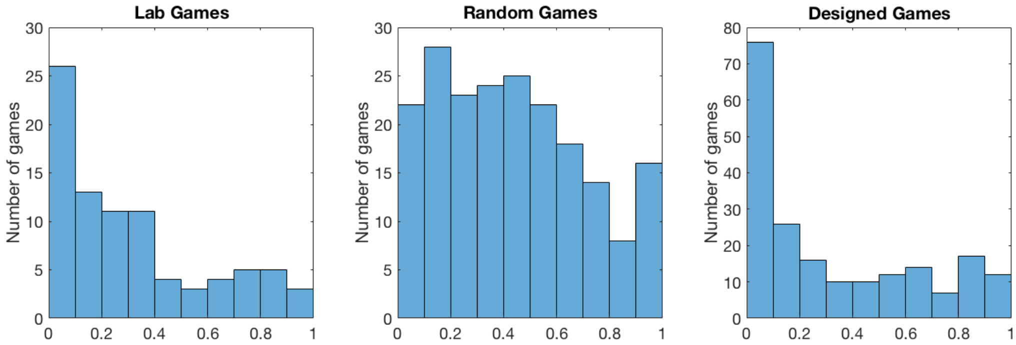


Figure 8: *Left:* Distribution of p-values across our set of lab games; *Center:* distribution of p-values across our set of randomly-generated games; *Right:* distribution of p-values across our set of algorithmically-generated games. For each set of games, the observed distribution of p-values is statistically different from uniform.

Thus, crowd predictions are at least in some cases drawn from a different distribution over actions than actual play. This suggests that it may be possible to improve upon the naive crowd rule by separating those games in which the crowd predicts well from those in which it predicts less well.[49]

7.3 Hybrid Models with Crowd Predictions

We thus turn to hybrid models that combine the crowd forecast with the models considered earlier: the level-1(α) model and PDNE.

[48] Our finding is similar in spirit to that of Costa-Gomes and Weizsacker (2007), who find (for a set of 14 lab games) that stated beliefs are closer to the uniform distribution than the actual distribution of play is.

[49] That the distribution of p-values is not uniform does not necessarily imply that there are games in which the crowd predicts better and games in which the crowd predicts worse. To take an extreme example, if all subjects always correctly predicted the modal action, the distribution of p-values would be far from uniform.

	Accuracy	Completeness
Guess at random	0.33	0%
Level-1(α)	0.68	52%
	(0.02)	
PDNE	0.56	34%
Crowd	0.76	64%
Level-1(α) + Crowd	0.76	64%
	(0.02)	
Crowd + PDNE	0.78	67%
	(0.02)	
Level-1(α) + PDNE	0.79	69%
	(0.03)	
Ideal prediction	1	100%

Table 9: Prediction accuracies for the hybrid models involving crowd forecasts.

The hybrid model that combines the crowd forecasts with PDNE performs about as well as the hybrid level-1(α) and PDNE model. We do not display its "model assignment tree"—the analog of Figure 6—because here the estimated tree varies too much from fold to fold. The model that combines level-1(α) with the crowd forecasts performs much less well. To understand the relative performance of the different hybrid models, it is useful to consider the correlations of their constituent models' predictions. The crowd predicts the level-1(α) action in 276 games (out of 486), so the predictive accuracies of these two approaches are highly correlated, as further detailed in the left table below:

Level-1(α) Crowd	Right	Wrong
Right	299	74
Wrong	24	89

PDNE Crowd	Right	Wrong
Right	198	175
Wrong	72	41

Table 10: *Left:* comparison of the crowd forecast and level-1(α). *Right:* comparison of the crowd forecast and PDNE.

There are only 24 games in which the level-1(α) prediction is correct while the crowd prediction is not. This greatly limits the potential of hybrid models combining crowd forecasts with level-1(α). Indeed, even if we learn a *perfect* assignment of games to models, the best achievable accuracy for the data set of all 486 games is $(486-89)/486 = 0.82$. In contrast, PDNE's prediction errors are far less correlated with the prediction errors of the crowd, which makes that hybrid more successful, just as having less correlated models is useful when building forecast combinations (Timmermann, 2006). When combining PDNE and the crowd predictions, a perfect assignment of games to models would attain accuracy of $(486-41)/486 = 0.92$; the fact that we only achieve accuracy of 0.78 with this hybrid shows there is scope for considerable improvement in our model assignment algorithm.

The correlation structure across model predictions also explains why the extension of our hybrid model to all three models (selecting whichever of PDNE, level-1(α), and the crowd forecast is predicted to perform best) does not improve on the PDNE-crowd hybrid.[50] In fact there are only 19 games (roughly 4% of the data set) in which the crowd prediction is correct, while both the PDNE prediction and the level-1(α) prediction are wrong.[51]

This small number of games is not enough for the addition of crowd predictions to our hybrid of level-1(α) and PDNE to result in better predictions. However, as in our exercise in Section 3.2, examining these games can help us identify features that the crowd seems to use but are not captured by either of those models.

One thing we observe is that the crowd outperforms level-1(α) and PDNE on games where some action is not part of a Nash equilibrium that isn't Pareto-dominant, but is nonetheless much more appealing than other equilibria, as in the game below:

	a_1	a_2	a_3	Frequency of Play
a_1	93, 93	10, 60	70, 53	53%
a_2	60, 10	30, 30	100, 33	40%
a_3	53, 70	33,100	10, 10	7%

Here, the crowd forecast correctly predicts action a_1. This action is part of a Nash equilibrium profile, but the corresponding payoffs $(93, 93)$ do not Pareto-dominate those of the two other pure-strategy Nash equilibria—$(33, 100)$ and $(100, 33)$. One way to capture this behavior may be to include a feature for whether there is a Nash equilibrium whose product of payoffs is "much larger" than that of any of the other Nash equilibria, or to compare the product of Nash equilibrium payoffs to those of all other action profiles.

Although our data has only a small number of games with this particular structure, we conjecture that with a data set that had a higher frequency of games like those above, we would find large improvements from crowd forecasts over level-1(α) and PDNE alone. If this is true, it would further reinforce the point that the performance ranking of different models depends on which games we examine. The mapping from games to behaviors or best-fit models, however, should remain fixed independently of how the experimenter samples across the space of games. Thus, better understanding of that mapping could be useful. These 19 games the crowd data identifies point the way to further improvements over the level-1(α) and PDNE models; we leave further exploration of such games to future work.

8 Conclusion

This paper uses approaches from machine learning algorithms not only to improve predictions of initial play, but also to improve our understanding of it. We use these tools to develop simple and portable improvements on existing models.

[50] The hybrid matches the performance of the best hybrid; both have an accuracy of 0.79.

[51] Here we set $\alpha = 0.41$, which is the median estimate from the set of all games across the different training sets.