2021

Deployment of Causal Effect Estimation in Live Games of Dota 2

Anders Harboell Christiansen

Emil Gensby

Bryan S. Weber
CUNY College of Staten Island

Deployment of Causal Effect Estimation in Live Games of Dota 2

Anders Harboell Christiansen
Independent Researcher
Lyngby, Denmark
andershc1993@gmail.com

Emil Gensby
Independent Researcher
Lyngby, Denmark
emil@gbcnet.dk

Bryan S. Weber
Department of Economics
College of Staten Island: CUNY
Staten Island, New York
bryan.weber@csi.cuny.edu

***Abstract*—In this paper, we provide an application that produces consistent in-game estimates of win probabilities in Dota 2. Previous work shows that common methods of identifying the effect of in-game features are strongly inconsistent, which we corroborate here with a large data set. We further provide an in-game application for players to see these estimates during the game as a training tool, along with displaying the estimated marginal impact of the primary actions (kills, last hits, and tower damage), which are previously known only by intuition. In a double-blind setting, we are the first to identify that users observe a difference between estimates produced by an inconsistent and consistent approach. Users show a significant preference for the consistent approach along several dimensions. Participants specifically identified the consistent approaches as having better quality advice by a large and significant margin, about four points on a ten-point scale.**

***Index Terms*—Instrumental Variables, Control Function, Causal Effects, Dota 2, Matchmaking, Winning Probability, User Experience**

I. Introduction

Players, spectators, and game developers have strong intuitions about the causal effect of player actions on the player's probability of winning. However, this intuition is not always perfectly aligned with the realities of the game. To this point, players have begun using Dota+ which provides win probability estimates for players [1].

Extending this example towards spectators, similarly, Weavr.tv obtained a £4 million grant for developing "immersive experiences and testing them with large scale audiences," and has provided in-game statistics and win probability estimates. At this time, it has been difficult to mobilize users and has taken a large investment to obtain less than 5,000 downloads [2]. Similar problems seemed to be present in Overwatch League Games - it was difficult to determine if the statistics had value and how to interpret them [3].

Our major contribution in this paper is to be the first, to our knowledge, to deploy a control function (CF) estimator for users in a live setting, and test it against a more traditional method. This application is available online here: [4]. We initially outlined the CF estimator in a previously published conference paper [5], as well as the related instrumental variables estimator. Here, we have created two versions of a live application, one using the CF estimator and the other using a parallel inconsistent estimator.[1] We then used those applications to conduct several surveys of users, finding that users saw the live CF application as useful. The CF estimates are rated as significantly "higher quality" than similarly calculated inconsistent estimates to a large degree (about a 4 point difference on a 10 point scale). To effectively deploy this application, we have greatly expanded our original data set of Dota 2 games from the OpenDota API [6].

The finding that live users seem to prefer causal effect estimates is particularly relevant to computer game analysis, where computer scientists are directed towards machine learning (ML) approaches by both training and familiarity [7]–[13]. The problem with this is that ML approaches emphasize precision in outcome prediction, as opposed to causal approaches which emphasize the consequences of manipulating a single factor while holding the remaining others constant [14].

In the remaining paper, Section II discusses the relevant literature on CF approaches and outcome

[1]An estimator is inconsistent if the estimate does not approach the value of the true population parameter as the sample size approaches infinity.

prediction in Dota 2. Section III articulates an operating definition for causal effects and how the approach differs from ML. Section IV covers the statistical approaches used and the nuances of this particular application. We then discuss the estimated causal effects and contrast them to the inconsistent estimators in Section V. These estimations (both inconsistent and consistent) are deployed in a double-blind manner in Section VI, and user surveys are analyzed. Section VII concludes with direction for future research.

II. Literature Review

When surveying the literature on games and win probabilities, we identified several areas of active research. We have found several papers which calculate the win probabilities in games of Dota 2. For example, [15] is based on the area between the players, their relative "inertia" towards the enemy base, and a few other positional components. Again, this paper does not claim that players win because of their inertia towards the enemy base, but rather, teams with strong inertia towards the enemy base tend to win. There can be other factors that trigger both inertia towards the enemy and victory, such as a surplus of in-game resources.

Several other papers have looked at predicting the outcome of Dota 2 games, however, all of these use logistic regression or other machine learning techniques. [7]–[13], [16]. These methods are inconsistent, meaning that while one may be able to, on average, predict the winner of a particular game, one cannot identify the change in the probability of winning after an outside update. For examples of an outside update, consider the consequences of new items being added to the game, or heroes being strengthened/weakened during balance updates.

In [5], we used the logistic method specifically as an example of an inconsistent method, though we emphasize inconsistency is present in numerous ML approaches, and highlight how to circumvent this problem. Here, we reiterate this problem, implement and release an in-game prediction method, and identify that participants find a large benefit to using consistent methods, a benefit which is significantly larger than that of an inconsistent one.

We also found patents based on calculating the probability of winning live games, for both sporting events [17] and slot machines [18]. The patents are broadly absent of technical details but point to a private incentive to capture probabilities and display them for users. The creators of Dota 2 sell such a service as part of Dota+ [1].

Last, we have also found some game balancing literature [19] which uses randomized Monte-Carlo actions to automatically determine the principal components of a game and the consequences of each of them. Since the Monte-Carlo actions are random and uncorrelated with previous actions, it seems plausible that this method can extract the causal effect of each action.

However, this approach does not appear to be plausible for games with arbitrarily large action spaces such as Dota 2. This Monte-Carlo simulation may also not take into account the distribution of actions that actual human players may explore, leaving a disconnect between the player's perspective of the value of a unit and the games' [20].

In this paper, we do not use randomization from automated Monte-Carlo experiments or synthetic players, instead we exploit a popular and voluntary randomization feature in the game (hero selection) in order to estimate the causal impact of particular game features on winning.

III. Causal Effects

Causal inference is the process of estimating how an independent alteration in a particular feature, holding all other factors constant, leads to a change in outcomes. We wish to contrast causation and correlation– both of which can make predictions. Formally, causation means that in an idealized randomized controlled experiment, we can expect an outcome Y as a result of a given intervention X. Correlation is simply a measure of the extent that X and Y occur together, even if the true cause of both lies elsewhere [21].

A person developing for game balance is, by definition, interested in the causal effect of the game features because they are performing an outside intervention in order to alter the game outcome. During a preliminary interview for this project, one game developer stated: "For example, if a player's win-rate changes significantly after buying a certain item in Dota 2 it could be a sign of a balancing problem." If players holding item X typically find they win the game, it is critical to identify if this is merely fashion (correlated) or if the item itself was causing the win

before manipulating the item's properties. This is trivial in the case of cosmetic items but becomes more complex when existing strategies and player habits center around particular items in a large ecosystem of equipment, heroes, and strategies.[2]

Directly, the problem of game balance is to compare the probability of winning, p, in two states: one state with the changes enabled, and one state without the changes. The difference in p between the two states is the causal effect of the treatment. Games where p is closer to 50% are presumably more balanced. Furthermore, changing a feature that is correlated with the game outcomes but does not change p in the expected direction or amount is problematic. For example, one would not wish to weaken a hero simply because they are new and the other players are unfamiliar with their abilities.

To estimate causation, most measures of fit are uninformative- such as R^2 which measures the predictive power of the features. Our focus is not on predicting the largest number of wins within our data set - we recognize there are many other inputs that could be included (ie, some researchers have replays and have scraped real-time data and item purchases). With these, the proportion of accurately predicted game outcomes could be increased [7], [12], [22]. Some researchers have also found techniques besides logistic regression obtain marginally higher predictive fits [7], but our emphasis is on the average partial effect of features [23]. We illustrate this problem in Probit estimation because of its commonality, but the problem persists in many settings: random forests [24], deep learning [25] and other commonplace ML techniques [14]. We utilize the CF approach to resolve this problem effectively.

IV. Methodology

A. The Domain of Dota 2

Defense of the Ancients 2 is a video game from 2013 that features several different competitive online five versus five modes. In the most common modes, players choose between a wide pool of heroes and they proceed to obtain gold and experience by killing units, called "last hitting" (LH). Additional resources can be obtained by killing enemy heroes and destroying enemy towers, but these two incomes are a secondary source of overall revenue. The win condition in Dota 2 is destroying the enemy's base, which is protected by a minimum of three towers. When one base falls, the other team wins.

Typically, after the matchmaking process has finished, all players select their unique hero from a diverse list. All heroes are listed with a primary attribute - strength (str), intelligence (int) or agility (agi) - which broadly indicates the function of the hero in the game (admittedly with exceptions). Generally speaking, str heroes have a lot of hit points but have less attack power. Heroes with int as their main attribute have access to immediately powerful abilities that do not scale very well into the late game, whereas agi heroes have the highest attack damage, particularly in the late game. A team of two str, two agi and one int could be an example of a balanced team. However, all random (AR) games in Dota 2 skip the selection process by randomly assigning a hero to all the players, much like an experiment. AR mode creates suboptimal teams - for example, a team of five int heroes would have a problem scaling into the late game (although there are some exceptions).

In our estimation process, we exploit this random assignment of heroes to identify the impact of game features on a team's probability of winning. This will be vital for providing consistent win probability estimations for the application and the experiment that follows.

We use AR games since All Pick or Captain's Mode do not have an easily exploitable random variation. Even better, since AR games have the same components as a standard game of Dota 2 (hero abilities, winning condition, etc.), AR games represent the minimum available deviation from the standard game. We wish to study typical player performance features such as last hits, kills, etc. Our aim is not to balance AR games – the mode is recreational and team balance is secondary. We focus on these AR games due to the randomizing aspect of hero selection that will strongly alter a team's in-game features.

This variation could not be caused by behavioral or strategic changes over the course of winning a game, since players have no choice in the assignment. This variation permits the estimation of causal impacts stemming from differences in the game's (presumably

[2]We note there is not an available corpus of Dota 2 games with randomized items or updates to item abilities, so we save this direction for future research.

random) starting hero composition and not from any other source. We then compare these estimated causal impacts and find them varying from the naive methods, and experiment to discover how players perceive the inconsistent and consistent estimates.

B. Data

Data were acquired as by issuing requests to the OpenDota API through Python [6], and appended to the existing data set [5]. The procedure for gathering data through the API is as follows:

1) Request a list of match IDs and their game mode.
2) Save the match IDs if the game mode tag is AR.
3) Request the match data for all the AR matches by using the match IDs.
4) Lastly, pull out the specified statistics from each match and insert them into a readable file that can be used for further experimentation.

Going forward we will be using the average per minute value for several gameplay statistics for each player. The statistics are as follows: average last hits per minute, average tower damage per minute, average kills per minute, the number of heroes belonging to each of the three primary attributes, the largest premade party size, and difference in matchmaking ranking (MMR) between the teams.

We note that the collected MMR is for competitive games in which the heroes are *chosen* by the player, unlike in AR games. To estimate the MMR of a team, we collected the rank of each player (when public) and used [26] to convert it to MMR. If a team had no public ranks listed we discarded the match, otherwise we used the average MMR for all public profiles on the team. We have collected a total of 5,155 matches with complete information, roughly doubling the size of [5] for a total of 10,310 observed sides. These data are summarized in Table I.

As we observe both sides of every match, the win rate of the teams in our data set must be exactly half. Broadly, these statistics remain consistent with our previous paper, [5], within 2 significant figures for all averages.

Kills are infrequent events, with each player averaging 0.183 kills per minute. Some particularly passive games resulted in 0 kills, where one side was unable or unwilling to kill any opponents. A back-of-the-envelope calculation suggests a player has a maximum of roughly 8 creeps per minute (the primary source of income), but each player averages around 3.6 last hits per minute.[3] The damage dealt to towers by each player averages to around 64 tower damage per minute. We calculated that players can win by taking a direct path to the enemy for as little as 7,800 total tower damage, and the maximum possible tower damage is 23,400 (excluding healing).

TABLE I: Data Summary: Player Averages for Both Teams

Statistic	Mean	St. Dev.	Min	Max
won	0.500	0.500	0	1
$kills/min$	0.183	0.071	0.000	0.873
lh/min	3.590	0.960	0.000	7.811
$towerdmg/min$	63.538	51.286	0	309.353
$largestpartysize$	2.558	1.090	1	5
str	1.625	1.028	0	5
agi	1.607	1.023	0	5
int	1.767	1.048	0	5
mmr_diff	0.000	747.634	−3,360	3,360
N	10,310			

AR mode is typically recreational and filled with premade parties (average premade team size of 2.6). AR mode in Dota 2 does not have a publicly available MMR, however, the game relies on the standard competitive Dota 2 skills, and so we reference the competitive MMR here. The distribution of str, int and agi is almost perfectly uniform among the three attributes, suggesting that the AR mode is indeed random.

C. Brief Overview

Broadly, the approaches we discuss in this paper are ways to exploit (truly) random variation. The random variation is used to experimentally "treat" different games, in our case by giving each team a different set of heroes. The approaches we enumerate below, in aggregate, examine the results of all of these small experiments across our data set. We also provide an outline of the theory demonstrating that naive approaches lead to inaccurate estimates of input effects.

[3]The number of creeps spawned increases as the game goes on, but 8/minute is the number of initially spawned creeps per lane.

D. Inconsistency in Simultaneous Equations

Let us consider a circumstance where one might accidentally obtain inconsistent estimates. Our example will center around *last hits* (LH), a critical measure of a team's net worth in Dota 2. LH is the primary means by which the team generates income and experience points for their heroes. The existence of a problematic feature itself is not unique to Dota 2; Starcraft will have similar problems using economic features [27], while first-person shooters like Halo will struggle with accuracy or kill/death ratios [28], [29]. We estimate a linear model for the sake of illustrating the issue:

$$win_i = \alpha_0 + \beta_1 LH_i + \epsilon_{1i} \tag{1}$$

However, the model does not have to be linear. The problem persists in various models and forms, including random forests [24], nonparametric functions [30], [31], and neural networks [25].

A researcher using the naive method is assumed to be interested in the causal effect of last hits on winning, estimated by the average partial effect: $\frac{dE(win_i|LH_i)}{dLH_i} = \beta_1$. Experience suggests players with many last hits become more powerful and eventually win, so we anticipate $\beta_1 > 0$.

But this is not the only relationship between LH_i and win_i. Players who are currently winning have the first-best choice of activities across the board and have opportunities to make last hits. At the same time, losing players must make due with second-best options such as using the jungle. Indeed, Dota 2 even has a mechanic which encourages damaging one's own units or towers to prevent last hits. This means one cannot reject the viability of Equation 2:

$$LH_i = \alpha_1 + \beta_2 win_i + \epsilon_{2i} \tag{2}$$

Here, we expect $\beta_2 > 0$, because winning provides numerous advantages which assist in getting last hits. A substitution exercise of Equation 1 into Equation 2 will find:

$$Cov(LH_i, \epsilon_{1i}) = \frac{\beta_2 \sigma_{1i}^2}{1 - \beta_2 \beta_1} \neq 0 \tag{3}$$

The existence of Equation 3 means one cannot directly estimate Equation 1. Estimating β_1 in Equation 1 will be inconsistent in the direction of $\frac{\beta_2 \sigma_{1i}^2}{1-\beta_2 \beta_1}$. Reiterating our earlier assumptions that $\beta_1 > 0, \beta_2 > 0$, we can show our estimates of β_1 will be too small if $Cov(LH_i, \epsilon_{1i}) < 0$, and too large if $Cov(LH_i, \epsilon_{1i}) > 0$ [21]. We point out that the magnitude of this inconsistency is often substantive, and in Section VI we verify that the inconsistency appears to be noticeable by players of various experience levels in a double-blind test, and seems to significantly degrade their opinions of the prediction in several ways.

This problem, called "endogeneity", is a major and reoccurring concern in econometrics literature [25], [32]–[35], one which is being reignited by the introduction of ML into the literature. An endogenous variable is influenced by either omitted variables, or by the outcome itself. Conversely, "exogenous" variables are independent and have no such confounding problem.

Control Function approaches (CF) (and Instrumental Variables (IV) estimators) allow us to calculate the causal effect of an increase in y_2 even if they are potentially endogenous. We outlined these approaches in [5], and wish to reemphasize that this problem is not exclusive to linear estimation, it spans a wide range of estimation techniques.

Formally, CF estimators calculate the average partial effect (APE) of an input on the predicted value of y_1. The APE is interpretable and tells why a team is winning rather than merely reporting the existing status, addressing some of the research goals of [36], [37]. Estimation of the APE requires: the endogenous y_2, exogenous x regressors, and instruments z, which must fulfill several requirements. These requirements are as follows:

1) Instruments must be uncorrelated with the error terms: $Cov(z, \epsilon) = 0$ (instrumental exogeneity)
2) Instruments must be a relevant predictor of y_2: $Cov(z, y_2) \neq 0$ (instrumental relevance)

To summarize, z *causes* x, but does *not cause* wins except when mediated through x [38].
The CF approach is a multi-step procedure for functions $f()$:

$$\begin{aligned} &a)\ y_2^{Stage1} = \beta_z^{Stage1} z + \beta_x^{Stage1} x + \epsilon_1^{Stage1} \\ &b)\ y_2 - \hat{y_2}^{Stage1} = \hat{\epsilon_1}^{Stage1} \\ &c)\ y_1 = f(\beta_x^{CF} x + \beta_y^{CF} y_2 + \beta_\epsilon^{CF} \hat{\epsilon_1}^{Stage1} + \epsilon_2^{CF}) \end{aligned} \tag{4}$$

See [5], [30], [31] for more details, but we note that each step is performed with the base package [39] in R. Proper standard errors from Equation 4c can be found by bootstrapping all 3 stages, as well as other methods

detailed in [38].

E. Challenges in Application & Practice

In the natural world, it is not guaranteed that viable instruments will exist. The instruments may be weak [40], or may remain endogenous despite a researcher's best efforts. Our instruments do appear to be perfectly exogenous, and we have highlighted several examples of even stronger instruments that meet all required conditions in [5].

Our instrument, however, has been found in the existing game of Dota 2. The AR mode is already available, which randomizes the heroes assigned to players and therefore the three primary attributes: $\{str, int, agi\}$. In other modes, we reiterate that the selection of heroes is not prior to the game [7], but a part of the game itself. Team composition is critical, though each hero's primary attribute is only a modest proportion of team composition, as recognized by [7], [8], [10], [11]. Having too many heroes with the same attribute can lead to a team missing certain capacities or abilities.

To provide an example of the importance of hero selection, OpenAI lost its only match when assigned a sub-optimal hero composition by the audience.[4] OpenAI gave itself a win probability of only 2.9% at the beginning of the game [41]. These randomly assigned hero attributes serve as experimental variation (instruments) for predicting those otherwise endogenous game features, and the combination of $\{str, int, agi\}$ has worked as a sufficient proxy in the past for this property. We note that there are fewer AR games available than other modes – but with a sufficiently large data set one could expand the set of instruments for each of the 121 unique heroes to address the challenge of each hero's uniqueness, as well as the nontrivial interaction effects between heroes [7], [8], [10], [11].

We repeat that the standard game modes as studied in [7], [10], [11] have teams alternate hero choices during the selection phase, making them endogenous – choice of hero depends on the other team and skill level of the players. By contrast, our sample focuses on heroes that are exogenously assigned prior to the game. There may be other remaining channels by which hero composition alters the probability of victory (such as particular item or hero synergies, etc.).

If those channels are small and the instruments (hero attributes) are strongly predictive, then any remaining inconsistency is small [21]. The exhaustive collection of controls as well as instrument choice is of utmost importance. Fortunately, feature collection is fairly comprehensive in this digital environment. We wish to highlight that perfect instruments can be inserted by game developers at will if the choice of instrument is still a concern.

The next step is to use these instruments in the CF framework and obtain estimates of the APE. We then find these consistent estimates and naive estimates, insert them into a live application, and display the predictions to the end application users. We then survey the users on their experiences.

F. Modeling with the Control Function Approach

The following three endogenous variables y_2 have been collected for both teams (marked in subscripts as a, b), where a represents the friendly team. All endogenous statistics are recorded as rates (per player per minute). We have collected six total endogenous variables, the same as collected in [5]:

1) lh_a, lh_b: The last hits, where a player kills an enemy non-hero unit. This is the primary method of collecting income.
2) $kills_a, kills_b$: The kills, where a player kills an enemy hero unit. This is the secondary method of collecting income.
3) $towerdmg_a, towerdmg_b$: The amount of tower damage dealt every minute. This is a tertiary method of collecting income.

The following four exogenous variables x are selected prior to the game and are not altered by any player decisions.

1) $mmrdiff_a$: The amount that team A is behind in MMR.
2) $largestpartysize_a, largestpartysize_b$: The number of players grouped together on the team.

[4]We note their composition had heroes that do not perform according to the stereotypical roles we listed above.

3) A vector of ones: To create an intercept term, included here to save on notation.

Last are the instrumental variables z. To repeat their three critical features: z are determined independently of player actions, z are expected to directly influence only the in-game statistics y_2, and there must be at least one z for each of the six variables in y_2. Here we use $C(attribute)_{team}$ to represent the count of heroes from $team$ with $attribute$ as factors.

1) $C(str)_a, C(int)_a$: This is a full set of indicator variables for *str* and *int* heroes that take the value 1 if team A has a particular count of that hero attribute or 0 otherwise. The counts range between 0 and 5, and must total to 5 or less.[5]
2) $C(str)_b, C(int)_b$: This is a second set of indicator variables for team B.

We intend to calculate the APE of y_2 on winning, win_a. In Section VI we display the APE to end users:

1) win_a: An indicator taking the value 1 if team A has won and 0 otherwise.

The CF estimator is the consistent analogue of Probit estimation (derived in [5]), where $\Phi()$ is the standard normal cumulative distribution function:

$$win_{ai} = \Phi(\beta_x^{CF-N} x_i + \beta_y^{CF-N} y_{2i} + \epsilon_{2i}^{CF-N}) \quad (5)$$

Holding the exogenous variables constant $\bar{x}$, we create point estimates of the APE at numerous values of last hits, kills, and tower damage. In Section V we contrast for readers the distinct partial effects shown by each estimator, and in Section VI we highlight that double-blind users prefer the consistent estimates.

V. Estimation

One might reasonably anticipate cross-validations at this point. One portion of the data is chosen as the training set and a model is fitted to this. This model is then fitted to a test set and R^2 or a similar measure is kept. Then the model is discarded. After multiple iterations, the measure of fits are compared and the best model is then fitted to the entire data set. This is not useful in this context, however, since obtaining better predictions is secondary to obtaining better estimated APEs.

We further highlight, along with [16], that our data comes from post-game statistics and therefore it is "very easy to find out who has won after someone has won a match." Our estimates are similarly precise, with the consistent CF model having an error rate of 0.023 and the inconsistent Probit model has an error rate of 0.024. The predictions of interest, using the games' randomized starting heroes to determine in-game statistics, are shown in Table II.

We contrast the APEs obtained by CF in Figures 1, 2, and 3 – pointing out they are distinct from the inconsistent estimates. In Section VI, we discuss the deployment and reception of an application that uses the consistent estimates developed in this section.

A. First Stage

The first stage array of tests are used for estimation of CF (as well as IV, see [5]). In the first stage, we estimate the endogenous variables, $lh_{a,b}, kills_{a,b}, towerdmg_{a,b}$, using the exogenous variables. The estimated residuals $\hat{y_2} - y_2 = \hat{\epsilon}$ are used in the CF. This eliminates any simultaneity problems as the course of the game turns to victory or defeat. We discuss the shared first stage and its predictive power in Table II.

TABLE II: Stage 1 Tests

	df1	df2	F-Statistic	P-value
Weak Instr. (lh_a)	20.00	10286.00	20.44	0.00^{***}
Weak Instr. ($kills_a$)	20.00	10286.00	4.76	0.00^{***}
Weak Instr. ($towerdmg_a$)	20.00	10286.00	3.85	0.00^{***}
Weak Instr. (lh_b)	20.00	10286.00	20.44	0.00^{***}
Weak Instr. ($kills_b$)	20.00	10286.00	4.76	0.00^{***}
Weak Instr. ($towerdmg_b$)	20.00	10286.00	3.85	0.00^{***}
Sargan	14.00		12.87	0.54
Note:				*p<0.1; **p<0.05; ***p<0.01

In this paper, we use the same instruments as in [5] for comparability. It is suggested to have an F-statistic over 10 [40], [42], implying not all of our instruments are strong but are particularly suitable for predicting last hits. Therefore, our estimates may have higher variance and may compound the errors between stages, creating bias in small samples, so we direct potential users towards created instruments as one way to avoid this problem. We attribute the remaining weakness in

[5]We note that $C(agi)_a = 5 - C(str)_a - C(int)_a$, so it is redundant to include $C(agi)_a$.

our instruments to the great variation in heroes, as well as the omission of synergistic effects between heroes. The larger data set suggests no evidence of over-identification [43] when tested by a Sargan test.

B. Nonlinear Control Function Approach

Next the naive Probit approach results are compared to our preferred specification, the improved CF approach. We note that this is a better estimate than the linear model because it accounts for diminishing marginal returns in the tails of the distribution and provides estimates that are properly bounded between 0 and 1 (representing loss and victory). The estimated APEs can be seen below, and the ± 1 standard deviations are obtained by bootstrapping 50 times. In our case, and in general, CF standard errors are fairly large, but have naturally improved as a result of increasing the size of the data set since [5]. We note that the APEs for team A and team B are simply inverses of one another, and so we will only show team A's estimated partial effects below.

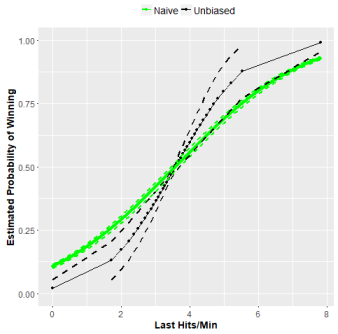


Fig. 1: Dotted lines represent ± 1 SD. The naive methods (green) puts less emphasis on last hits than the CF approach (black). All else being equal, there is a positive association between winning and last hitting.

Figure 1 displays the anticipated positive association between last hits, the primary method of strengthening a hero and winning. The inconsistent estimates (green) are extremely narrow relative to the consistent CF (black), a feature common to all of the inconsistent estimates in Section V-B. Despite these error bounds, there is an appreciable dissimilarity between the CF approach and the naive approach, larger than 1 SD in either direction.

Evaluating Figure 1 assuming an average player has 5 last hits per minute (holding other inputs constant at the mean), the naive estimates suggest they are about 69% likely to win the game whereas the CF estimates suggest that the team is 80% likely to win. To replicate this inconstancy, one might assert that winning teams emphasize damaging towers and pushing rather than performing additional last hits. The CF approach removes this inconsistency by exploiting random deviations when the team roles are not filled.

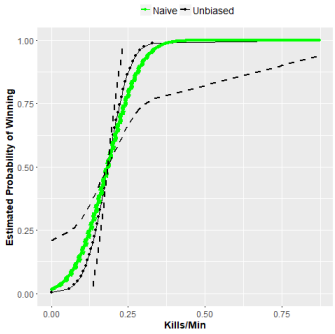


Fig. 2: Dotted lines represent ± 1 SD. The CF approach (black) leaves kills relatively unchanged. Additional kills remain an important predictor of victory.

Figure 2 shows the estimated probability of winning given a particular level of kills, holding all other inputs constant at their respective means. Again, the inconsistent naive estimates (green) have extremely narrow standard deviations. These inconsistent estimates suggest that a team with players 1 SD below the mean, at about 0.1 kills per minute, are 16% likely to win.

The CF estimates (black) suggest that an equivalent team is even less likely to win (8%), although the standard error bounds are fairly comprehensive and no significant difference is visible between the two. All else being equal, both estimates agree that if a team has an average of approximately 0.37 kills per minute, (2.6 standard deviations above the average) then victory is nearly certain.

Figure 3 shows the effect of the final major input on the probability of winning, tower damage per minute. Identically to the previous figures, the inconsistent naive estimates (green) have a narrow confidence interval. All else being equal, we evaluate the naive estimates at about 1 standard deviation above the average, approx. 100 tower damage per minute, and find it suggests that the team is almost certain to win (94%).

Intuitively, this estimate seems overstated, and the absolute pinpoint precision of the estimator seems problematic rather than reassuring. On the other hand,

the CF estimates (black) suggest a team with equal damage per minute remains reasonably matched with their opponents (55%).

We believe this is because winning teams must eventually destroy towers (since it is the mechanism by which teams win the game). Therefore, standard techniques have trouble distinguishing between the conditions that permit winning (setting up powerful heroes in a coordinated attack) and the actual act of winning. As a result, the green approach conflates the two. Compared with the smaller data set in [5], this point estimate has improved by reversing to appropriately represent the direction of the effect, though the SD remains wide.

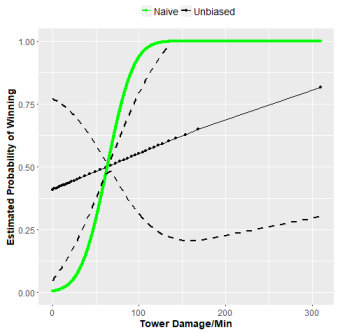


Fig. 3: Dotted lines represent ± 1 SD. The CF approach (black) suggests the amount of tower damage a team deals has only a modest effect on the game. The inconsistent estimates (green) suggest that even a small amount of additional tower damage will easily decide the game.

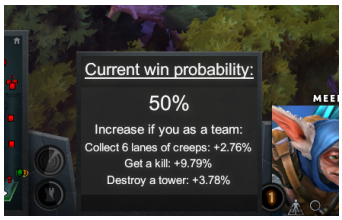
We believe this overstatement (or understatement) of the inconsistent estimators APE, particularly those shown in Figures 1 and 3 is noticeable and creates the difference in user experience shown in Section VI.

We also emphasize it is important to have a sufficiently large data set for analysis, highlighting that one of our coefficients has changed sign to a more sensible direction after doubling the size of the data set. To reiterate, we can explain the difference between the inconsistent and consistent estimates is that choosing "first-best options" is the prerogative of the winning teams – they end the game by attacking towers. Conversely, losing teams are forced into "second-best options" – defensive positions that tend to avoid attacking enemy towers – collecting last hits and perhaps even trading tower hit points for security.

VI. Deployment and Results

Unique to this paper, we have implemented the previous results as an application for Dota 2. We named this application "Winnr" and it is available for public downloads on the Steam Workshop [4]. The application was created using the Dota 2 modification tool "Hammer" which is available through Dota 2 itself on Steam. The win percentage for the team in question is then displayed in the lower-left corner next to the in-game mini-map. Images of the application itself can be seen in Figure 4.

(a) Full Screen View



(b) Relevant Winnr Addition

Fig. 4: Winnr Screenshots

In the application we track the teams' last hits, kills and tower damage, averaged over their respective teams such that we can feed the statistics to the win probability calculation along with the game duration. Since the model is designed to predict win or loss using per minute inputs averaged over the full game, and at the early game there are so few minutes played, we noticed the application's predictions fluctuated dramatically in the early game.

To adapt, we decided to hardcode the game time input to the median game time length of about 30 minutes, or the actual game time, whichever is larger. This means that early on our model tends towards 50% in the first part of the game, but it becomes more decisive the longer the match is. Additionally, because our estimates are taken from post-game statistics, this estimate provides the probability of winning given that the players were to conclude the game at that particular moment.

As an addition to the overall win probability calculation, we decided to add three suggestions that would help players optimize and increase their chances of winning. The three suggestions are tied to the marginal effect of the three input variables: last hits, kills, and tower damage. This serves as a basic training application and helps players weigh their decisions about primary game activities.

The first suggestion is based on last hits, in which the team is told how their probability of winning will change if they collect six lanes of creep spawns without incident.[6]
The second suggestion is based on kills, in which the team is told what happens to their win probability if they get an additional kill on an enemy player without incident.
The third suggestion is based on tower kills, in which the team is told what happens to their win probability if the team manages to destroy one of the opposing towers without incident.

The marginal win probability is calculated by evaluating the game state x_{state} and counterfactual x'_{state} and comparing $f_{CF}(x'_{state}, t+1) - f_{CF}(x_{state}, t)$, where f_{CF} is the control function approach calculated above. As an example, for estimating the marginal effect of kills, the counterfactual state has the prevailing number of kills + 1.

In order to examine if this application was desirable, we preliminarily examined quantitative survey results from the Foundations of Digital Games (FDG) conference, and through Mechanical Turk.[7] We managed to collect eight respondents from these sources, four from Mechanical Turk and four from FDG. We named this group the *Informed CF survey group*, since we were directly collecting results to see how individuals responded to the application and participants were informed of the application's contents. Their responses are given in Table III and discussed below.[8]

[6]The variable amount of creeps that spawn throughout the game is taken into account as the creep waves become larger as the game progresses.

[7]Our requirements on Mechanical Turk were that users have made online purchases of video games, and their primary internet device is desktop.

[8]One individual took the survey more than once, so we have averaged that individual's responses.

We then improved this design to use a double-blind procedure and collected 12 more participants from numerous online sources, bringing our total participation to 20 respondents. To elicit participation, four participants, selected at random from among complete respondents, were offered $25 via PayPal after the survey concluded.

We solicited participants from numerous subreddits (Dota2, LearnDota2, TrueDota2, SampleSize, Playtesters, Playmygame), Dota 2 and game-related Discord groups, Twitter, LinkedIn, Facebook groups and the TeamLiquid forums. Participants were randomly assigned either the naive model (control) or the improved CF model (treatment) we have evaluated here in Section V. This assignment was done by consulting a random number generator within Dota 2 at game start. This random number altered the model displayed and, after the game, displayed a link taking them to the appropriate post-game survey.

We note that a median game of Dota 2 takes 30 minutes to complete, so some individuals may not have completed the entire game, but we do not have a record of them or their survey results. These double-blind results are shown together with the informed CF survey group results in Table III.

TABLE III: Quantitative Survey Results

		Double-Blind	
	Informed CF Survey Group	Inconsistent Model	Consistent CF Model
Category	Mean	Mean	Mean
Experience Level	$5.23^{\dagger\dagger\dagger}$	6.00	7.25
Perceived Accuracy	8.05	6.00	6.13
Quality of Advice	6.19	$3.25^{\dagger\dagger\dagger}$	7.00^{***}
Desire for Integration	$6.29^{\dagger\dagger}$	$7.00^{\dagger}$	8.25^{*}
New Players Benefit	$6.57^{\dagger\dagger}$	$7.00^{\dagger}$	8.75^{*}
Recommend to Friend	$6.28^{\dagger\dagger}$	$5.00^{\dagger\dagger}$	8.13^{**}
Number of Participants	8	4	8

Note: 10 is high for all categories, 1 is low for all categories.
***,**,* = Significantly different from double-blind, inconsistent model at a 1%, 5%, 10% level.
†††,††,† = Significantly different from the double-blind, consistent model at a 1%, 5%, 10% level.

The informed CF survey group rated themselves

generally less experienced than the participants we managed to recruit for the double-blind experiment. This difference in experience was significant between the informed survey group and the consistent CF group, suggesting perhaps our advertised venues were targeting a higher level of player.[8] The consistent CF group generally rated the application positively, rating all aspects of the program as over 5 on average, including the anticipated benefit to new players. The consistent CF group had no significant differences from the informed survey group, though the informed participants were introduced to the application in person and perhaps were less interested because there was no compensation.

In the double-blind experiment, the consistent CF group ranked the experience as better than the inconsistent model group in every single category, and these differences are typically significant. When presented with double-blind exposure between naive and CF approach, participants in the consistent CF group rated their advice as higher quality by nearly four points on a ten-point scale. We consider this result the most important one and it is also the largest magnitude of any difference in our survey results - it indicates that the results are not just theoretically better as discussed in [5], but we now provide evidence that players themselves rate predictions made by the CF model as higher quality.

VII. Conclusion

In this paper, our contribution is to highlight that individual players rate consistent estimators as having significantly higher quality of advice (4 points on a 10 point scale), in a double-blind test. Participants also rate the consistent estimators as more desirable for integration, more beneficial to new players, and more likely to recommend to a friend. This has been done by creating an entirely new application, available on the Steam Workshop, called "Winnr" for Dota 2 that utilizes these improved estimates. We, to our knowledge, are the first to show a preference for consistent estimators via experiment.

Previous work in [5] has shown that consistent estimators have numerous advantages over the alternatives. In particular, the point estimates typically are in a more sensible direction and have a more reasonable magnitude. We have also improved our estimates from [5] by doubling the size of the data set, and as a result one of our potentially concerning coefficient estimates from [5] has now adjusted to the proper direction. Using the greatly expanded data set from OpenDota [6], we highlight that the naive approach is inconsistent by a substantial amount, overestimating the win probability by nearly 40% in relatively common game-states. We speculate that these distinctions in forecast outcomes are the root of player preferences for consistent estimates, since they were not informed of any other characteristics of the estimation process.

Relevant for practitioners, the naive estimates also provide far too confident estimates of the causal effects, though these confidence intervals were not shown to players in the application. These inconsistencies are a consequence of winning teams being able to perform certain first-best activities while losing teams tend to be forced into second-best options. As a result, naive methods aggressively associate the second-best options with losing, and the first-best options have overstated benefits.

Both of these results suggest that game developers, players, and spectators should be cautious when utilizing inconsistent estimates to predict the impact of game modifications, in-game estimates of win probabilities, or the importance of a particular action in spectator software. As demonstrated in this paper, individuals should either reference consistent estimates or at least be aware of the appropriate interpretation of inconsistent estimators.

For future work, instrumentation has not yet been exploited in relation to player position or item selection. This might be a route to explore given concerns about advising systems causing more rapid convergence towards a particular meta, as pointed out by [44].

Appendix

A. Survey

1) What is your experience level with Dota 2 (or similar games)?
   Scale: 1 (no experience) to 10 (I am in the top 1% of players).
2) I feel the tool's forecasts of win probability were accurate.
   Scale: 1 (strongly disagree) to 10 (strongly agree).
3) I feel this tool typically gives good advice.
   Scale: 1 (strongly disagree) to 10 (strongly agree).
4) I would like an option for this tool to be integrated into the game.
   Scale: 1 (strongly disagree) to 10 (strongly agree).

5) I believe new players would benefit from this tool. Scale: 1 (strongly disagree) to 10 (strongly agree).
6) I would recommend this tool to a friend. Scale: 1 (strongly disagree) to 10 (strongly agree).
7) In which scenarios would you use this tool?
8) Do you have any other thoughts about Winnr?

REFERENCES

[1] Dota Team, "Introducing dota plus," Online., Mar. 2018. [Online]. Available: http://blog.dota2.com/2018/03/introducing-dota-plus/

[2] Weavr, "Weavr dota 2 companion - apps on google play," Nov. 2020. [Online]. Available: https://play.google.com/store/apps/details?id=com.Rewind.WEAVR&hl=en_US&gl=US

[3] "Broadcasters struggling to give context and meaning to game stats - by karahol," Dec 2017. [Online]. Available: https://www.winstonslab.com/news/2017/12/07/owl-overwatch-context-statistics/

[4] A. H. Christiansen, E. Gensby, and B. S. Weber, "Winnr," Dec. 2020. [Online]. Available: https://steamcommunity.com/sharedfiles/filedetails/?id=1942027992

[5] ——, "Resolving simultaneity bias: Using features to estimate causal effects in competitive games," in *2019 IEEE Conference on Games (CoG)*. IEEE, 2019.

[6] OpenDota, "Dota 2 statistics," Mar. 2019. [Online]. Available: https://www.opendota.com/

[7] Y. Yang, T. Qin, and Y.-H. Lei, "Real-time esports match result prediction," *arXiv*, Dec 2016.

[8] A. Semenov, P. Romov, S. Korolev, D. Yashkov, and K. Neklyudov, "Performance of machine learning algorithms in predicting game outcome from drafts in dota 2," in *International Conference on Analysis of Images, Social Networks and Texts*. Springer, 2016, pp. 26–37.

[9] I. Makarov, D. Savostyanov, B. Litvyakov, and D. I. Ignatov, "Predicting winning team and probabilistic ratings in "dota 2" and "counter-strike: Global offensive" video games," in *International Conference on Analysis of Images, Social Networks and Texts*. Springer, 2017, pp. 183–196.

[10] N. Wang, L. Li, L. Xiao, G. Yang, and Y. Zhou, "Outcome prediction of dota2 using machine learning methods," in *Proceedings of 2018 International Conference on Mathematics and Artificial Intelligence*, 2018, pp. 61–67.

[11] K. Song, T. Zhang, and C. Ma, "Predicting the winning side of dota2," Stanford University, Tech. Rep., 2015.

[12] V. J. Hodge, S. Devlin, N. Sephton, F. Block, P. I. Cowling, and A. Drachen, "Win prediction in multiplayer esports: Live professional match prediction," *IEEE Transactions on Games*, vol. 13, no. 4, pp. 368–379, 2021.

[13] F. Johansson and J. Wikström, "Result prediction by mining replays in dota 2," Master's thesis, Blekinge Institute of Technology, 2015.

[14] J. Pearl, *Causality*. Cambridge university press, 2009.

[15] F. Rioult, J.-P. Métivier, B. Helleu, N. Scelles, and C. Durand, "Mining tracks of competitive video games," *AASRI procedia*, vol. 8, pp. 82–87, 2014.

[16] L. Kinkade, N. Jolla and K. Lim, "Dota 2 win prediction," University of California, San Diego, Tech. Rep., 2015.

[17] M. Kerns, J. Ma, S. McClelland, and M. Kamal, "Win probability based on historic analysis," Mar. 29 2007, uS Patent App. 11/606,970.

[18] B. Schultz, "Gaming device and method with bonus and displayed winning probabilities," Mar. 18 2004, uS Patent App. 10/245,550.

[19] P. Beau and S. Bakkes, "Automated game balancing of asymmetric video games," in *2016 IEEE conference on computational intelligence and games (CIG)*. IEEE, 2016.

[20] M.-V. Aponte, G. Levieux, and S. Natkin, "Measuring the level of difficulty in single player video games," *Entertainment Computing*, vol. 2, no. 4, pp. 205–213, 2011.

[21] J. M. Wooldridge, *Introductory econometrics: A modern approach*. United States, Boston: Cengage, 2015.

[22] P. Yang, B. E. Harrison, and D. L. Roberts, "Identifying patterns in combat that are predictive of success in moba games." in *FDG*, 2014.

[23] F. Johansson, U. Shalit, and D. Sontag, "Learning representations for counterfactual inference," in *International conference on machine learning*, 2016, pp. 3020–3029.

[24] S. Athey, J. Tibshirani, and S. Wager, "Generalized random forests," *arXiv*, 2018.

[25] M. H. Farrell, T. Liang, and S. Misra, "Deep neural networks for estimation and inference," *Econometrica*, vol. 89, no. 1, pp. 181–213, 2021.

[26] Dota 2 Wiki, "Matchmaking/Seasonal Rankings," Mar. 2019. [Online]. Available: https://dota2.gamepedia.com/Matchmaking/Seasonal_Rankings

[27] B. S. Weber, "Standard economic models in nonstandard settings–starcraft: Brood war," in *2018 IEEE Conference on Computational Intelligence and Games (CIG)*. IEEE, 2018, pp. 417–424.

[28] D. Buckley, K. Chen, and J. Knowles, "Rapid skill capture in a first-person shooter," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 9, no. 1, pp. 63–75, 2017.

[29] K. J. Shim, K.-W. Hsu, S. Damania, C. DeLong, and J. Srivastava, "An exploratory study of player and team performance in multiplayer first-person-shooter games," in *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*. IEEE, 2011, pp. 617–620.

[30] R. Blundell and J. L. Powell, "Endogeneity in nonparametric and semiparametric regression models," Tech. Rep. CWP09/01, 2001.

[31] R. W. Blundell and J. L. Powell, "Endogeneity in semiparametric binary response models," *The Review of Economic Studies*, vol. 71, no. 3, pp. 655–679, 2004.

[32] J. D. Angrist and A. B. Keueger, "Does compulsory school attendance affect schooling and earnings?" *The Quarterly Journal of Economics*, vol. 106, no. 4, pp. 979–1014, 1991.

[33] D. Card, "Using geographic variation in college proximity to estimate the return to schooling," Tech. Rep. 4483.

[34] C. M. Hoxby, "Does competition among public schools benefit students and taxpayers?" *American Economic Review*, vol. 90, no. 5, pp. 1209–1238, 2000.

[35] W. N. Evans and R. M. Schwab, "Finishing high school and starting college: Do catholic schools make a difference?" *The Quarterly Journal of Economics*, vol. 110, no. 4, pp. 941–974, 1995.

[36] Z. Yang, Z. Pan, Y. Wang, D. Cai, X. Liu, S. Shi, and S.-L. Huang, "Interpretable real-time win prediction for honor of kings, a popular mobile moba esport," *arXiv*, 2021.

[37] Z. Yang, Y. Wang, P. Li, S. Lin, S. Shi, and S.-L. Huang, "Predicting events in moba games: Dataset, attribution, and evaluation," *arXiv*, 2020.

[38] J. M. Wooldridge, *Econometric analysis of cross section and panel data*. MIT press, 2010.

[39] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2017. [Online]. Available: https://www.R-project.org/

[40] J. Bound, D. A. Jaeger, and R. M. Baker, "Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak," *Journal of the American statistical association*, vol. 90, no. 430, pp. 443–450, 1995.

[41] OpenAI, "Openai five benchmark: Results," Mar. 2019. [Online]. Available: https://www.opendota.com/

[42] D. Staiger and J. H. Stock, "Instrumental variables regression with weak instruments," National Bureau of Economic Research, Working Paper 151, January 1994. [Online]. Available: http://www.nber.org/papers/t0151

[43] J. D. Sargan, "The estimation of economic relationships using instrumental variables," *Econometrica*, vol. 26, 1958.

[44] A. De Venecia, "Dota plus subscription: Is it worth it (review and guide)," Aug. 2020. [Online]. Available: https://robots.net/gaming/dota-plus-subscription-is-it-worth-it-review-and-guide/