Measuring Randomness in Tabletop Games

James Goodman, Diego Perez-Liebana, Simon Lucas
Game AI Research Group
Queen Mary University of London
james.goodman, diego.perez, simon.lucas@qmul.ac.uk

***Abstract*—Tabletop games often incorporate random elements in the form of dice or shuffled card decks. This randomness is a key contributor to the player experience and the variety of game situations encountered. There is often a tension between a level of randomness that makes the game interesting, and a level at which the outcome itself is effectively random and the game becomes dull. The sweet-spot for any given game will depend on the design goals and target audience. We introduce a new technique to quantify the level of randomness in game outcome due to these elements. We use this to compare 15 different tabletop games, and then to disentangle the different contributions to the overall randomness from specific parts of the game. We show the utility of this approach by using it with a game publisher as a tool in the development phase of a new commercial board game.**

I. Introduction

Different games have different levels of randomness. A perfect information game like Chess has none. Poker or Bridge have random deals of cards to players so that each hand is different; and the fact that players' cards are hidden from the other players adds additional uncertainty to the game outcome.

All of these examples have been very popular games over a period of decades through centuries, so it is clear that having random elements in a game does not make it a bad game, nor does their exclusion. In modern game design in general, and in modern tabletop games in particular, the use of random outcomes through dice, shuffled card decks or similar devices is a central part of the design palette [1]–[3]. Elias et al. 2012 note a number of reasons for indeterminacy to be included, covering both directly random events and the role of hidden information. These include variety of game-play (as in the random deal in each hand of Bridge), extending the range of interesting competitors (so a family game may have higher levels so that a young child can still win reasonably often without detracting from the game experience), and adding interest to outcomes as the player await the reveal of a new hole card in Poker, or the result of a doubling throw in Backgammon [2]. Adding too much of the 'wrong' sort of indeterminacy can destroy a game for a target audience. An extreme case being Snakes and Ladders, in which the outcome is entirely unaffected by player skill.

As Elias et al. note, measuring the randomness in a game is not trivial. A game with a single dice throw that decides the outcome has one random event, while one that has 100 such dice throws may be much less random in terms of outcome as these may effectively average out. Also of importance is how player skill interacts with randomness to mitigate bad luck, and efficiently exploit good luck.

This paper looks at the effect of stochasticity in a number of different tabletop board games to quantify the impact of randomness in each game. The questions asked are:

1) "Does the randomness in the game affect the result?" Specifically, do different initial shuffles of the cards, or predetermined runs of dice results, gives an advantage to one player, and if so, how big is this effect?
2) If this randomness does have a significant effect, can the different contributions to this be disentangled? For example in Seven Wonders there is a deck of 'Wonders', with each player receiving one randomly at the start of the game. There are separately three other Decks of cards, one for each of the the three rounds of the game. How much of the overall effect is due to each of these sources.

These questions are of interest in game design. In some cases specific sources of randomness will be desirable, and others not - all depending on the design goals. We apply the technique introduced to a game currently in commercial development as an illustration of how this can provide concrete data to support designer intuitions and drive changes to the final design.

II. Background

A. TAG and Games

The games analysed are all commercially tabletop board or card games published in the last 50 years. They are implemented on the Tabletop Games Framework (TAG) for ease of comparison [4]. TAG is a research framework that facilitates comparative research on these games. The code is publicly accessible at https://www.tabletopgames.ai/Resources.html.

The detail of the 15 games included in the overall analysis (see Section IV) is omitted for space and summaries of them are available at www.boardgamegeek.com. Three games are analysed in more detail in Section IV-A, and the relevant aspects of these are described below:

- **Seven Wonders** is a set collection game in which a final tableau of 18 cards is scored by each player. Each player is dealt one of 7 'Wonder' boards at the start. This provides player-specific resources, abilities and/or victory point options. Each player is then dealt a hand of 7 cards, all players play one simultaneously, then pass the remaining 6 to the

979-8-3503-5067-8/24/$31.00 ©2024 IEEE

player to their left. This continues until each player has played 6 cards (the last is discarded). A new deck ('Age II') is then dealt, and the process repeated. Finally a third deck ('Age III') is dealt and played. Each player can interact with the two neighbouring players, and there are hence interactions between different boards. In a 3-player game there are 70 possible set ups (the clockwise order matters).

- **Colt Express** is a game about a train robbery, in which players must plan actions into an uncertain future, and taking into account what the other player may be doing. Each player is dealt one of 6 characters, each with a special ability. The train to be robbed is then set up by dealing N+1 random train tiles (N = number of players), and the round structure determined by dealing 5 random round cards (only the top one is ever visible). Each round card defines how the players play cards to rob the train, and possibly an event that occurs at the end of the round affecting all players. Each player has their own deck of action cards, which is shuffled and they draw a hand of 5; the discard pile of action cards is reshuffled to form the new deck as needed. Over 5 rounds the players play action cards and seek to maximise their score from robbing the train. Given all combinations of characters, round and train cards in a 2-player game, there are over 2 million setups possible, before considering the shuffling of the individual player decks.
- **Dominion** is a deck-building game. Each player starts with a deck of 10 cards - 7 Copper (money) and 3 Estates (victory points). These are shuffled and the player draws 5 cards as their hand. On their turn a player can play an Action card (if they have one), and buy a card from one of 10 possible Action cards, or another money/victory card. When a player exhausts their deck, they shuffle their discard pile to renew it. Money and Action cards are more useful at the start of the game to build an engine that generates victory points in late game.

B. MCTS

Monte Carlo Tree Search (MCTS) has been very successful in many games [5], [6]. It has been adapted to imperfect information games with Information Set MCTS used in this paper [7].

The root node is initialised with the current state. MCTS proceeds by repeating four stages every iteration until the available computational budget is exhausted, at which point it returns the action with the best estimated value.

- **Selection**. The next action at a node is the one with highest Upper Confidence Bound (UCB) on the expected value. This points to the next node in the tree. This is deterministic and balances exploitation of empirically good actions with exploration of those tried less frequently [5]. This continues until a node is reached with previously untried actions.
- **Expansion**. If the current node has untried actions, then one is picked at random and a new node added to the tree for the state that is reached from taking this action.
- **Rollout**. After expanding a new node, actions are taken by some defined rollout policy (the default being random actions) until the game end is reached and a win/loss or score result obtained.
- **Back-propagation**. The result of the rollout is then back-propagated up the tree through all nodes that were passed through in the Selection and Expansion phases. The statistics on each node are updated to track the number of times each action was taken on previous iterations and the average win/loss or score that this resulted in. These statistics then drive choices in future Selection phases.

For the purposes of the discussion in Section IV, the main points to note are that although the selection phase is deterministic *given* previous iterations, both Expansion and Rollout phases make random choices and hence for any specific random seed MCTS may recommend different actions for the same number of iterations from the same root state.

C. Player skill and game stochasticity

Amy and Boris are playing Chess. Amy is better than Boris, but will she win every game? Classic approaches to skill measurement assume not, and instead assume that the probability of victory is skewed towards Amy, but that this is not guaranteed. Chess is a deterministic game, so this variability in outcome must instead come in from the players themselves. Perhaps Amy decides to play the King's Gambit opening, which Boris is very familiar with. If she had instead decided to try the King's Indian Attack, then he would have been clueless and put up little resistance. At a more tactical level, human players make mistakes. On Turn 14 Boris may forget about the latent fork on his Queen as he enthusiastically takes the offensive, despite having spent time thinking about the possibility on Turn 9. The net result is that between two *human* players, the result of the game is probabilistic, even though the game itself is deterministic with perfect information available to both players.

Now consider two AI players, Alpha and Beta. If these algorithms are themselves deterministic, such as minimax search to different fixed depths, or using different state evaluation heuristics, then we expect every game of Chess they play to give the same result. In fact every game they play will have exactly the same sequence of moves. In this case, we cannot measure the skill difference between the agents using the same method.

If however these players are algorithmically stochastic, then we get back to a more 'human'-like environment. This is the case in Monte Carlo Tree Search for example. While each decision in the tree is deterministic using the UCB formula, the search is still stochastic from two inputs:

1) Random order of expansion of untried actions at a node;
2) Random actions taken during rollout.

This assumes different random seeds are used for the agents in each game; if the same seed is used each time, then we hit the same issue of a fixed outcome as with the purely deterministic agents. Hence using MCTS agents, provided they do not all use the same random seed, facilitates the approach of measuring the win rate over 1000 games in which the random seed is fixed.

III. Previous Work

Selecting specific seeds in games to define the setup is a standard tactic and is core in Perfect Information Monte Carlo, in which N possible samples of an Imperfect Information game are each solved using minimax search or other technique, and then the results amalgamated to make a decision in the real, unknown game [8], [9].

It can also be used to provide a sample of games as 'easy' or 'difficult' for AI agents [10]. This also holds for human play, for example in Duplicate Bridge identical deals of cards are played by all tournament participants so that their skill can be fairly compared. A pair may 'lose' a hand, but gain tournament points because in relative terms they lost less than other pairs [11].

In tabletop games randomness is a core design tool as outlined in Section I, and this is discussed in detail by Kaufeld 2011 [12].

Different levels of randomness are appropriate for different audiences and target game experiences. A useful distinction can be made between 'randomness' and 'uncertainty'. The deal of cards gives every player *random* hand, but even if others know what cards they have they will be *uncertain* about which they have played face-down on the table. See [13] for a discussion, but this work deals purely with the pure *random* effects. In many cases a game can be wildly 'unfair' due to this randomness and still meet its design goals [14], [15].

In other cases this issue can become detrimental to a game with more experienced players, who may seek more 'balanced' versions. One example of this in the games used here is in Catan, where the initial random board layout can give a significant benefit to the first players to pick the location for their starting settlement [16], This does not stop Catan being a very successful game with over 40 million copies sold and a competitive World Championship series [17].

Varying the seed used by the MCTS player has also been used to provide a number of functionally different players, despite using exactly the same algorithm [18], [19]. Analysis of this in 2-player perfect information games shows that the win rate of individual MCTS seeds can vary quite dramatically in line with the discussion in Section II-C, and the distribution of this variance can be plotted [19], [20]. This distributional consideration is perhaps closest to this work, but we look at Game seeds instead of Player Seeds.

IV. Methodology

For each game 100 different random seeds are sampled. For each random seed 1000 games are run and the win rate of the first player measured (a draw is counted as 0.5 of a win). For each of these 1000 games the MCTS players have different random seeds. This gives a sample of 100 different win rates. If the stochastic elements of the game have no net impact on the outcome then we expect this distribution of 100 win rates to be tightly clustered around X%. If the game is perfectly balanced with no first player advantage then $X = 50\%$, but this is not required. Each of the mean win rates is the average of 1000 independent games, so under the assumption of a binomial distribution with N = 1000 and p = X, 99% confidence bounds can be calculated. We expect on average 1 of the 100 win rates to fall outside these bounds.

The metrics assessed for each game are:

1) Entropy of the distribution of win rates. The win rates are discretised into 2% buckets, and the Shannon entropy, S, of this discrete distribution calculated, $S = \sum_i -p_i \log p_i \; i \in 1..50$.
2) The number of samples that fall outside the 99% binomial confidence interval. If this is much larger than 1 then there is strong evidence that the outcome of the game is affected by the specific random seed.
3) Span of win rate. The maximum win rate of the 100 samples minus the minimum win rate. This will vary from 0 (all game seed give the same win rate), to 1.0 (seeds can vary from a 100% win rate for the first player to a 0% win rate).
4) Trimmed Span. To reduce sensitivity to outliers, take the central 95% of the Span, discarding the most extreme 5% of samples. (See Figure 2 for a graphical example.)

All games are run for the minimum number of players they support. This is either 3-players (Hearts, Seven Wonders, Catan, Puerto Rico), or 2-players for the other eleven games. For all experiments MCTS agents with a 50ms computational budget per decision are used. The MCTS parameters are tuned separately for each game to ensure that the agent plays reasonably well. For details on these parameters see [21].

The distinct random seeds are:

1) A seed used to control all game events (deck shuffles, dice throws etc). This is akin to the chance player in OpenSpiel [22][1].
2) A seed used to redeterminise the state to hide hidden information before being passed to the MCTS agent for a decision. This needs to be kept distinct from the game chance player. For example in Poker a single hand has a variable number of actions. If no players bid beyond the mandatory blinds, then there is one decision per player. If any players bid then there are additional decisions. At each of these decisions the game state is redeterminised before being passed to the agent for a decision. If this process used the main game seed then the shuffle of the deck for the next hand would change depending on how many actions had been taken previously. We require, for a fixed game seed, the same sequence of future shuffles regardless of the player decisions.
3) A seed for each player used within the MCTS algorithm described in II-B. Importantly the random seed of the game is not known to the player, who makes decisions based purely on the current public information.

It is vitally important to ensure that there is a clear separation of the random seeds that drive game actions, and the random seeds that drive agent decision making via MCTS, and also that these random seeds drive all the variability

[1] https://openspiel.readthedocs.io/en/latest/concepts.html

of outcome. As preparatory work for each game this was confirmed by running several games with fixed random seeds for the game and for all agents and confirming that each game played out identically.

A. Disentangling contributions to randomness

For some games, such as Poker or Hearts, there is no source of randomness apart from the shuffle of a single deck of cards at the start of the game, or at the start of each round.

In others, there is a natural separation of 'sources' of randomness. In Seven Wonders every player is dealt a Wonder board at the start of the game from a shuffled deck. This provides them with unique options, and their game is also affected by the Wonders of the other players; players can buy resources from their immediate neighbours and some Wonders are more geared to specific strategies. If only one Wonder in play has a focus on 'Science' cards, then this player will have an advantage vis-a-vis an alternate deal in which several players have Wonders with this focus. The other source of randomness comes from the three decks of Age cards, which are each shuffled once at the start of the game.

For some games additional random seeds were added to the implementations to measure the impact of these different sources. For example, in Seven Wonders a seed was added to control the shuffle of the Wonder boards, and a separate one to control the shuffle of the Age cards. Two sets of experiments are then run, repeating the previous methodology, i.e.

- Sample 100 seeds for the Wonder board shuffle (or the Age deck shuffle)
- For each of these seeds run 1000 games with all other sources of randomness (game or player) initialised differently for each game.

The games selected for this approach are:

- Seven Wonders. A separate seed for the Wonder board shuffle is introduced, and one for all the Age decks (i.e. a single seed is used by all three decks, not one each).
- Colt Express. Three new seeds are introduced to control each of:
  - the shuffle and deal of each player's character;
  - the shuffle of the cards that make up the train carriages;
  - the shuffle of the round cards.
- Dominion. One new seed is introduced that controls the initial shuffle of the starting deck of ten cards. This determines the first two hands a player receives, which will either split the 7 Copper cards 3/4 (5 times in 6) or 2/5 (1 in 6). The first two turns of a player in Dominion and their choice of purchased card can be vitally important. This tests how important this effect is.
- Catan. An initial intention was to have a distinct seed for the map set up in Catan, to separate the effect of this from the random dice rolls and shuffle of the development card deck. This is in line with the observations on the impact this can have on a game's outcome [16] . However, MCTS agents with a 50ms budget are very poor at Catan given the very large branching factor and long game length. Given the results in Section V-A, this intention was dropped.

B. A commercial game

Working with Bright Eye Games[2], a commercial board game publisher, this approach has been applied to a game currently in the development process with a working title of 'Theme Park'.

The game is set in a theme park, with a number of children being taken on a day out and all wanting to go on different rides. The children will have damaging tantrums if they do not get what they want. The winner is the player who best balances their competing demands by closing time. The game has two key decks of cards:

- A set of Person cards, of which each player receives four at the start of the game. These are public and define which rides each child wants to visit.
- A set of 'Magic' cards that are one-off special abilities. These are face-down until drawn and placed on the map, at which point a player can take an action to pick them up for later use.

There are no other sources of randomness or hidden information in the game One question posed during the development process is whether one or other of these two decks is too 'random' and unbalancing. An answer to this question was sought by separating random seeds for each of these two sources, and applying the technique in Section IV-A.

V. Results

A. Overall

Figure 1 plots the histograms of the first player win rates for each seed. If these all fall within the green-bounded 99% confidence interval for a game then the random seed has no impact on the outcome. Encouragingly this is the case for the two purely deterministic games included as controls, Connect 4 and Dots and Boxes. It is also true for Diamant (aka Incan Gold). This is a push-your-luck game with a random deal of cards to form a cave system to explore. Players act simultaneously and have no distinguishing abilities or cards, and hence this result is as expected. Each game will play out differently due to the random deal of cards but each player has an equal win chance regardless of the deal.

The impact on outcome is distinct from the first player advantage that exists in some games. This is observed when the green confidence interval is to the right of the red bar in Figure 1. For example in Can't Stop (mean first player win rate is 55%), where each player acts in turn and the game ends as soon as one player hits the victory condition, potentially giving the first player an extra turn. Similarly, Dots and Boxes has a small first player disadvantage, although the game is perfectly 'fair' in the sense that luck has no effect on the outcome.

Catan is well known to be affected by the random set up of the board, but this is not evident in Figure 1 [16]. MCTS with only a 50ms budget and no game-specific heuristics is

[2]https://www.brighteyegames.com/

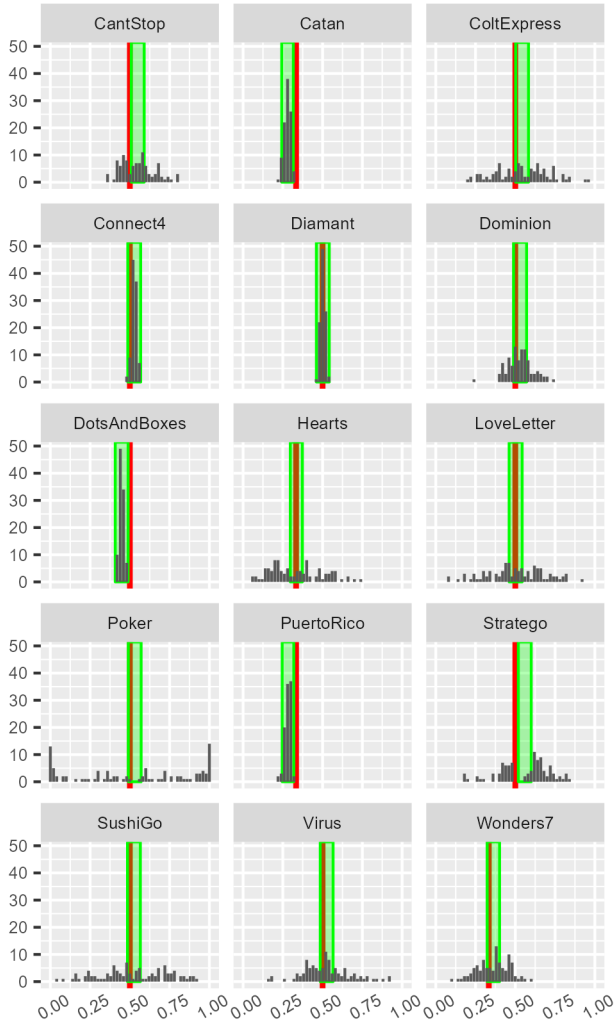

Fig. 1. Random Seed plot for all 15 games. The red line marks the win rate for the first player of 50% for 2-players and 33% for 3-players (if no first-player advantage). The green shading is the 99% confidence interval for the win rate assuming that the random seed has no effect. The x-axis is first player win rate, and the win rates for the 100 seeds are plotted as a histogram with buckets of width 2%.

very poor at the game, and this seems to lie behind the lack of any variation in outcome across different random seeds. Using better agents/heuristics in this more complicated game is anticipated to change the results here in future work. This may also be an issue with the Puerto Rico results.

Table I summarises the metrics for the experiments represented in Figure 1. Span, Entropy and Outliers all concur on the general pattern of which games have an outcome highly dependent on the specific random seed.

The entropy of the distribution can be less informative as the examples of Sushi Go and Poker make clear. Sushi Go has the highest formal entropy because the distribution of mean outcomes has the flattest overall distribution. Poker has a lower entropy, but this is because there are two peaks in the distribution at 0% and 100%, with the Span and even Trimmed

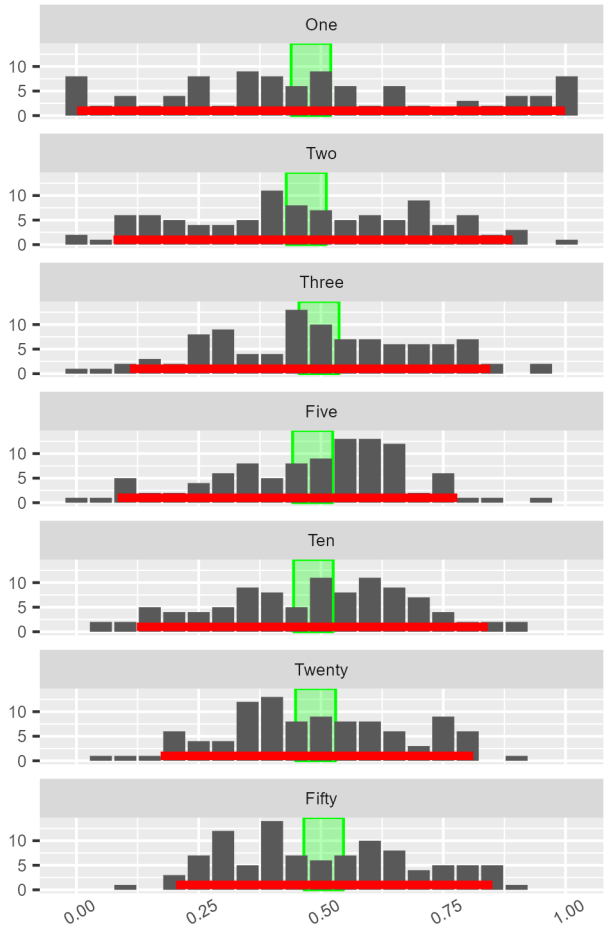

Fig. 2. Plots of random seed effects as the number of rounds to win in Love Letter is increased. For 1 round the shuffle often means one player is predetermined to win. As more rounds are required the effect reduces. Buckets of width 5% are used to better illustrate the pattern. The thick red line above the x-axis marks the Trimmed Span in each case. This is 1.0 for 1 round at the top, with its lowest value of 0.64 at the bottom for 50 rounds.

Game	Players	Span	T-Span	Entropy	Outliers
Dots and Boxes	2	0.06	0.06	1.13	0
Diamant	2	0.08	0.05	1.16	1
Connect 4	2	0.08	0.06	1.21	2
Catan	3	0.09	0.07	1.44	3
Puerto Rico	3	0.10	0.08	1.32	5
Dominion	2	0.51	0.29	2.65	55
Can't Stop	2	0.45	0.40	2.84	68
Seven Wonders	3	0.48	0.36	2.92	73
Virus	2	0.76	0.64	3.18	74
Love Letter	2	0.83	0.66	3.35	82
Colt Express	2	0.75	0.60	3.30	83
Sushi Go	2	0.89	0.74	3.49	83
Hearts	3	0.67	0.59	3.22	89
Stratego	2	0.66	0.63	3.02	94
Poker	2	1.00	1.00	3.26	97

TABLE I

SPAN, TRIMMED SPAN, ENTROPY AND OUTLIER METRICS FOR EACH GAME. SPAN IS THE DIFFERENCE BETWEEN THE FIRST PLAYER WIN RATES IN THE BEST AND WORST SEEDS; TRIMMED SPAN REDUCES THE IMPACT OF OUTLIERS, AND REMOVES THE 5% MOST EXTREME SEEDS FIRST. ENTROPY IS THE ENTROPY OF THE HISTOGRAMS IN FIGURE 1. OUTLIERS ARE THE NUMBER OF SEEDS OUTSIDE THE GREEN 99% CONFIDENCE INTERVAL IN FIGURE 1.

Span at the maximum possible value of 1.0, and 97 of 100 seeds being outside the 99% confidence interval. For some shuffles of the deck it is simply impossible to (fail to) win. This is with a modicum of skill so that good cards are taken advantage of; this does *not* mean that a random agent would win in these cases.

The implementation of Poker in TAG gives each player 50 chips, with a big blind of 10. Hence each game only lasts for a short number of hands (often just one if player's go All-In). This also explains the high impact of the random seed.

Figure 2 illustrates the same effect in Love Letter of increasing the number of points (i.e. rounds) needed to win the game. The standard rules for the 2-player game require a player to win 7 rounds (and more than the opponent). If this is amended to a single round, then each game will last a single round as one player must win each, and the pattern in Figure 2 is similar to Poker, with many deals fundamentally predetermined to be won by one player or the other. As the number of required points increases this effect declines slowly, potentially allowing the desired level of overall randomness in outcome to be tuned.

Stratego is a deterministic hidden information game, but without any random shuffling or dice rolls. It is therefore expected to have a pattern similar to Dots and Boxes or Connect 4, with the distribution mostly within the green confidence interval. The deviation from this expectation in Figure 1 is due to the detail of the implementation in TAG. Stratego starts with a setup phase in which the two player decide where to place their pieces on the board; the other player knows where the pieces are but not their identity. In TAG this phase is skipped, and the pieces of the two players are allocated randomly to one of three predefined setups. This gives nine distinct setups, the play of each of which then has no in-game stochasticity. The results for Stratego are therefore actually 9 distinct clusters, and Figure 1 shows that some of these are more favourable to one or other player.

B. Disentangling sources of randomness

Figure 3 and Table II summarise the experiments to disentangle the contribution of different sources of randomness.

For Seven Wonders the random distribution of the player boards has more impact on the game outcome than the shuffling of the three Age decks, although both are significant contributors to the total variation. Holding only the board seed constant leads to 55 of 100 tournaments being outside the 99% theoretical confidence bounds; 39 of 100 are outliers when the card seed is fixed.

For Colt Express the impact on the game outcome is driven by the initial deal of each player's character (66 outliers), with much smaller contributions from the Train (7) and Round (19) card decks. In Dominion there is a similar prominence in the importance of the initial shuffle of the player decks. Once this is taken into account, the later deck shuffles have much less effect.

The player boards in Seven Wonders and the players' characters in Colt Express are public information. This can

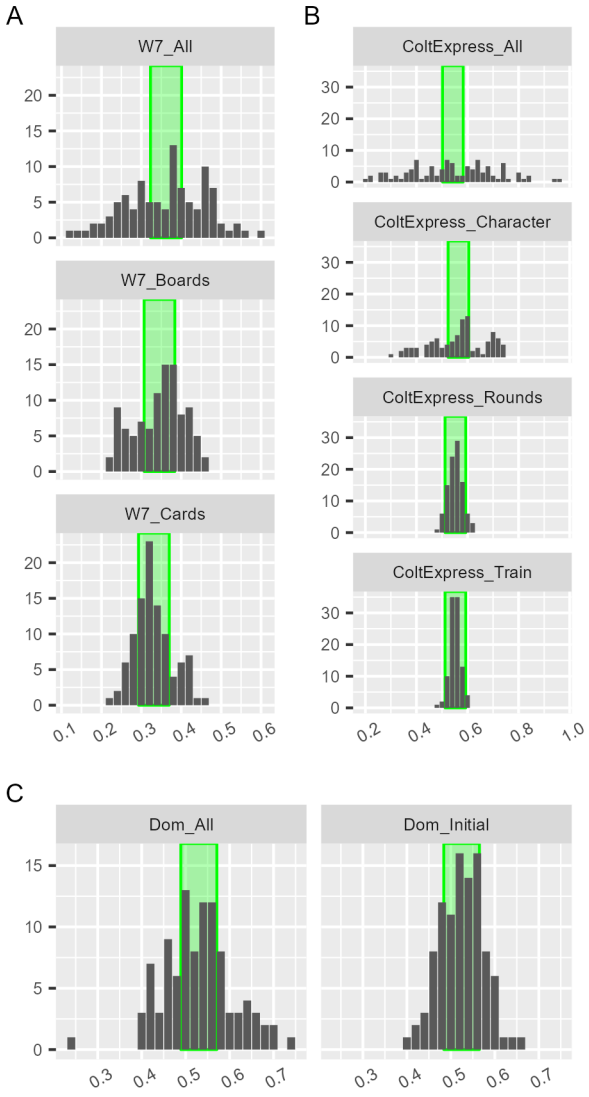

Fig. 3. Comparison of components of randomness. A) For Seven Wonders; All is the full results (As in Figure 1, Boards is for 100 fixed wonder board set ups (with Cards then varying); Cards for 100 fixed initial Card Shuffles (with Boards varying). B) For Colt Express breaks out the impact of initial Character, Round and Train shuffles. C) For Dominion showing the impact of just the initial shuffle and distribution of starting cards.

lead to experienced players knowing before the game starts which players have a positional advantage, evidenced by player discussions for both games on forums [23], [24].

The designer of Colt Express weighed in on this discussion, and clarified that perfect balance was not a key criterion, and subsidiary to making each character unique and interesting, "Colt Express is a fun game and indeed balance issues were not my first focus. I wanted to get a fun game and I'm quite happy with the result. In Colt Express, I wanted everyone to have a fun time and not only the guy who will win. I wanted to give each character a strong identity. Each ability is unique to make each player want to play at least one time

Game	Constant	Span	T-Span	Entropy	Outliers
Seven Wonders	All	0.48	0.36	2.92	73
Seven Wonders	Wonder Board	0.24	0.21	2.43	55
Seven Wonders	Age Cards	0.26	0.18	2.23	39
Dominion	-	0.51	0.29	2.65	55
Dominion	Initial Shuffle	0.24	0.18	2.30	43
Colt Express	All	0.75	0.60	3.30	83
Colt Express	Character	0.43	0.38	2.86	66
Colt Express	Train	0.11	0.08	1.48	7
Colt Express	Rounds	0.14	0.11	1.77	19
Theme Park	All	0.56	0.44	2.94	76
Theme Park	Person Cards	0.37	0.29	2.61	54
Theme Park	Magic Cards	0.43	0.31	2.73	71

TABLE II
RESULTS FOR DISTRIBUTION OF WIN RATES FOR THE FIRST PLAYER ACROSS 100 DIFFERENT RANDOM SEEDS FOR EACH GAME AND ANALYSED SOURCE OF RANDOMNESS. 'CONSTANT' INDICATES WHAT IS HELD CONSTANT WITHIN EACH OF THE 100 SAMPLED SEEDS; 'ALL' CORRESPONDS TO THE DATA IN TABLE I.

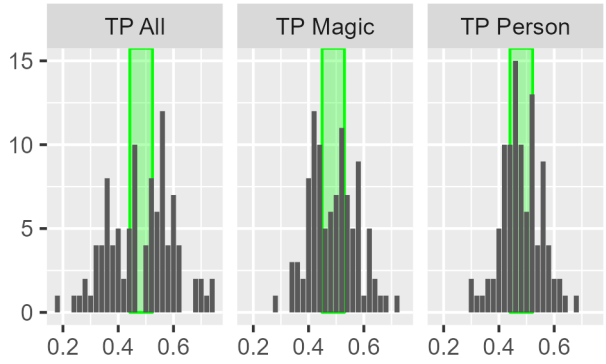

Fig. 4. Plot of win rates for first player for 100 random seeds for Theme Park (left), then fixing just the initial shuffle of the Magic deck (centre) and the Person deck (right).

each character of the game. Each ability needs you to play each character differently. So the conclusion is: I had to create quite unbalanced abilities." [23]

The random outcome in Dominion is dominated by the initial shuffle and the content of the first two hands. Fixing just this first shuffle still gives 43 outliers, and a trimmed span range of win rate of 18%. Fixing *shuffles* (the 'All' column) has a trimmed span of 29% and 55 outliers. A game of Dominion using the recommended first game cards usually involves 5-10 reshuffles of their deck after the original one.

C. A commercial game

The results for Theme Park are shown in Figure 4 and Table II. In this case the total variation in outcome is more evenly balanced between the two components, with the Magic Card deck shuffle having a slightly larger impact; 71 outliers and a trimmed span of 31% win rate variation versus 54 and 29% for Person cards.

The Person cards are all dealt at the start of the game and are public information. In this they are similar to the boards in Seven Wonders and the characters in Colt Express. Unlike those two games the Person cards do not have unique abilities and flavour, and each is just a list of 3 different rides the Person is keen to go on. The variation in outcome stems from synergies (or their lack) between the four cards that a player receives. The same ride appearing on 2 or 3 different cards, or rides across different cards being near to each other, makes it easier to visit the rides in the time available.

The Magic cards in contrast are all unique special abilities, and are only revealed during play for players to pick up and use. This means that any inbuilt advantage to one player is less visible at the start, and much of the fun of the game comes from strong and varied effects of these cards.

As a result of this analysis, the publisher decided to look into options to reduce the variation arising from Person cards by having predefined sets of starting cards, each of which could be more balanced. They were much less concerned about the variation from Magic cards.

VI. DISCUSSION

Tables I and II show that there are significant differences between games when we measure the mean win rate of the first player for a fixed random seed. This is not of itself 'good' or 'bad', and depends on the design goals. In the case of Colt Express the designer has stated that the impact of the random allocation of characters is deliberate, or at least that balance is a low priority compared to flavour and varied player experience, and a similar point has been made by the designers of other board games [14].

Measuring the overall randomness of a game is distinct from identifying 'strong' and 'weak' cards or characters. One unbalanced character may strongly affect a small subset of games for example without affecting the holistic measure much. Additionally this holistic measure of randomness across the game captures interactions between different elements, where the strength of a card depends on what other cards are in play.

In Theme Park, these results have helped prompt a re-design of one aspect of the game as having too strong a role for relatively undifferentiated player set ups was not wanted, while the slightly larger effect of the Magic card shuffle was acceptable.

All the various metrics tried show the same pattern across games. The Entropy measure has been least useful as it gives lower values for a game with peaks at 0% or 100% win rates, such as Poker or 1-round Love Letter, over games with narrower span but more central distribution. Both the Trimmed Span and Outlier count give a more robustly useful single number. In practice when working with game designers, the graphical displays of the distributions are much more interpretable, and are a richer information source.

There are a number of weaknesses and opportunities for further development of this approach. A primary weakness is that the agents used have had a low computational budget of 50ms per decision. In most of the games this is an adequate player, but one that can be beaten by intermediate humans as represented by the authors. In the more complex games,

such as Catan and Puerto Rico, this represents very poor play. Repeating these experiments with a better agent in these games is feasible, but has a high compute cost with 100,000 games required. A related issue is that only MCTS agents have been used, albeit with their parameters tuned to each game considered. Other techniques may be more suited to some of these games, and it is possible that using a counterfactual-regret based approach could change the result of games like Poker [25].

Plans for future work are to consider:

- Variation in ability. The agents used have deliberately been identical in terms of ability. It will be interesting if the same patterns occur for different skill levels - both for homogeneous agents (does greater skill increase or decrease the ability to mitigate stochasticity) and heterogeneous (if player 1 is better than player 2, does that reduce the random effect). Are there games for which randomness has a large effect on beginner players, but not for more skilled players?
- Error bound calculation. The wide variation in outcomes for different random seeds has an impact on reported error bounds on the win rate of a game. If a game is unaffected by stochasticity, then the classic use of a binomial model with a fixed p (win rate) to calculate these is fair. However, if each individual game played has an actual p that varies widely around a mean p' due to the random seed then this means more games need to be run for the same confidence interval on the result. Calculating the correction required is current work in progress.
- Seed-picking. This work has just looked at the overall distributional effect of random seeds. Curating individual seeds for games that have strongly skewed results can provide insight to a designer as to the detailed causes. Curating seeds with a balanced win rate can be helpful to better measure the relative performance of two agents, and a sequence of increasingly 'hard' seeds could be used to build a curriculum for training reinforcement learning agents.

VII. Conclusion

We have presented a new technique to quantify the inherent randomness in a game as measured by the distribution of first player win rates for a set of 100 different random seeds with equally skilled agents. We have conducted a comparative analysis of 15 different popular tabletop board and card games, and for three of them analysed the contribution of specific aspects of the game. The insights these provide can be of use to the game designer as we show with a small case study using the technique in a game currently in development by a commercial publisher.

This technique is of general use as a tool for game design as well as for the analysis of existing games. A game is not 'bad' if randomness has a big impact on the game outcome (or vice versa). What matters is whether the level of impact is in line with the design objectives for the game.

References

[1] J. Schell, *The Art of Game Design: A book of lenses*. CRC press, 2008.

[2] G. S. Elias, R. Garfield, and K. R. Gutschera, *Characteristics of games*. MIT Press, 2012.

[3] S. Woods, *Eurogames: The design, culture and play of modern European board games*. McFarland, 2012.

[4] R. D. Gaina, M. Balla, A. Dockhorn, R. Montoliu, and D. Perez-Liebana, "TAG: A Tabletop Games Framework," in *Proceedings of the AIIDE workshop on Experimental AI in Games*, 2020.

[5] L. Kocsis and C. Szepesvári, "Bandit based monte-carlo planning," in *European conference on machine learning*. Springer, 2006, pp. 282–293.

[6] C. B. Browne, E. Powley, D. Whitehouse, S. M. Lucas, P. I. Cowling, P. Rohlfshagen, S. Tavener, D. Perez, S. Samothrakis, and S. Colton, "A Survey of Monte Carlo Tree Search Methods," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 4, no. 1, pp. 1–43, Mar. 2012. [Online]. Available: http://ieeexplore.ieee.org/document/6145622/

[7] P. I. Cowling, E. J. Powley, and D. Whitehouse, "Information Set Monte Carlo Tree Search," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 4, no. 2, pp. 120–143, Jun. 2012. [Online]. Available: http://ieeexplore.ieee.org/document/6203567/

[8] M. L. Ginsberg, "GIB: Imperfect Information in a Computationally Challenging Game," *Journal of Artificial Intelligence Research*, vol. 14, pp. 303–358, Jun. 2001. [Online]. Available: https://jair.org/index.php/jair/article/view/10279

[9] D. Rebstock, C. Solinas, M. Buro, and N. R. Sturtevant, "Policy Based Inference in Trick-Taking Card Games," *arXiv preprint arXiv:1905.10911*, 2019.

[10] K. Sfikas and A. Liapis, "Collaborative Agent Gameplay in the Pandemic Board Game," in *FDG '20*, 2020.

[11] W. Hosch, "Duplicate Bridge," 2006. [Online]. Available: https://www.britannica.com/topic/bridge-card-game/Duplicate-and-tournament-bridge

[12] J. Kaufeld, "Randomness, Player Choice and Player Experience," in *Tabletop Analog Game Design*, ser. Costikyan, G., & Davidson, D.(Eds.). ETC Press, 2011.

[13] G. Costikyan, *Uncertainty in games*. Mit Press, 2013.

[14] P. Olotka, "Fair Isn't Funny: The Design of Cosmic Encounter," in *Tabletop Analog Game Design*, ser. Costikyan, G., & Davidson, D.(Eds.). ETC Press, 2011.

[15] F. Reiber, "Major Developments in the Evolution of Tabletop Game Design," in *Proceedings of the 3rd IEEE Conference on Games*, 2021.

[16] Schreiber, "Settlers of Catan," in *Tabletop Analog Game Design*, ser. Costikyan, G., & Davidson, D.(Eds.). ETC Press, 2011.

[17] K. Teuber, "About Us | CATAN," 1995. [Online]. Available: https://www.catan.com/about-us

[18] W. M. Czarnecki, G. Gidel, B. Tracey, K. Tuyls, S. Omidshafiei, D. Balduzzi, and M. Jaderberg, "Real World Games Look Like Spinning Tops," *arXiv preprint arXiv:2004.09468*, 2020.

[19] D. L. St-Pierre and O. Teytaud, "The Nash and the bandit approaches for adversarial portfolios," in *2014 IEEE Conference on Computational Intelligence and Games*. IEEE, 2014, pp. 1–7.

[20] J. Liu, O. Teytaud, and T. Cazenave, "Fast seed-learning algorithms for games," in *Computers and Games: 9th International Conference, CG 2016, Leiden, The Netherlands, June 29–July 1, 2016, Revised Selected Papers 9*. Springer, 2016, pp. 58–70.

[21] J. Goodman, D. Perez Liebana, and S. Lucas, "Skill Depth in Tabletop Board Games," in *IEE Conference on Games 2024*, 2024.

[22] M. Lanctot, E. Lockhart, J.-B. Lespiau, V. Zambaldi, S. Upadhyay, J. Pérolat, S. Srinivasan, F. Timbers, K. Tuyls, and S. Omidshafiei, "OpenSpiel: A framework for reinforcement learning in games," *arXiv preprint arXiv:1908.09453*, 2019.

[23] C. Raimbault, "Colt Express balance," 2016. [Online]. Available: https://boardgamegeek.com/thread/1440249/article/33201263#33201263

[24] J. Clark, "The wonders are significantly unbalanced," 2021. [Online]. Available: https://boardgamegeek.com/thread/2697126/wonders-are-significantly-unbalanced

[25] M. Lanctot, V. Lisỳ, and M. Bowling, "Search in Imperfect Information Games using Online Monte Carlo Counterfactual Regret Minimization," in *AAAI Workshop on Computer Poker and Imperfect Information*, 2014.