The Cross-sectional "Gambler's Fallacy": Set Representativeness in Lottery Number Choices

Jaimie Lien*
Department of Economics
School of Economics and Management
Tsinghua University

Jia Yuan*
Department of Business Economics
Faculty of Business Administration
University of Macau

Current Version: October 7th, 2014
Initial Version: October 31st, 2013

Abstract:[1]

Traditionally, the Gambler's Fallacy is described as the belief that a sequence of independent outcomes over time should exhibit short-run reversals. The underlying psychological bias thought to drive this fallacy is Representativeness Bias: the idea that even a small sample of outcomes should closely reflect the theoretical probability distribution (Tversky and Kahneman, 1971). Yet representativeness also has less commonly explored consequences in the cross-sectional dimension. We find strong evidence for this in lottery play where probabilities are well-defined and transparent, using a dataset of over 1.6 million lottery tickets purchased by over 28,000 players. Specifically, individuals prefer number combinations that are cross-sectionally representative of the uniform distribution from which they are drawn. We test two possible approaches to implementing representativeness; a heuristic 3-bin approach which is promoted in some gambling advice literature, and a direct optimization approach in which gamblers try to spread the numbers in the chosen set as evenly as possible across the lottery number range. By both measures, gamblers over-gravitated to highly representative lottery number sets and over-avoided less representative sets, compared to the proportions that the true lottery odds would suggest. In this pari-mutuel lottery setting, a cost is incurred by gamblers with this type of bias, by reducing their expected winnings.

Keywords: Belief Biases, Representativeness Bias, Gambler's Fallacy, Lottery Gambling
JEL Codes: D01, D03, D81, L86

* Corresponding Authors: jaimie.lien.tsinghuasem@gmail.com; jiayuan@umac.mo ; We are grateful to Vincent Crawford, Tao Li, Guang-zhen Sun, and Jie Zheng for helpful comments and conversations. We are indebted to our editor William Neilson for valuable advice which has improved the paper. Jaimie Lien acknowledges financial support from the National Science Foundation of China (#71303127), the Ministry of Education (China), and Tsinghua University. Jia Yuan acknowledges financial support from the University of Macau (# MYRG046-FBA12-YJ). All errors are our own.

1. Introduction

The traditional Gambler's Fallacy is usually exemplified by the following situation: Suppose that a fair coin is to be flipped several times in sequence. The coin has been flipped three times so far with the outcome HHH ("H" denoting heads, and "T" denoting tails). The Gambler's Fallacy predicts that if you ask individuals to guess the outcome of the next flip, they are likely to think the next flip will be T instead of H, even though both outcomes have equal probability under the fair coin assumption. The underlying reason for the fallacy is thought to be Representativeness Bias (Tversky and Kahneman, 1971), also known as the Law of Small Numbers (Tversky and Kahneman, 1971; Rabin, 2002). Decision-makers believe that the short sequence of coin flip outcomes should reflect the actual 50-50 probability distribution, leading them to believe that T is due to appear in the next flip.

In this paper, we show that another consequence of Representativeness Bias appears in cross-sectional decisions, not requiring a sequential process. If decision-makers believe that a small sample of outcomes from a particular theoretical distribution should closely reflect that distribution; when they are asked to pick a *set* of outcomes, they believe that set should "look" like the theoretical distribution. This is a new dimension of representativeness that has not been rigorously analyzed before, is readily illustrated by the case of lottery number choices. Consider a lottery game in which 6 integers from the range 1 through 33 will be chosen by randomly drawing 6 labeled balls from a cage containing the 33 labeled balls. This scenario is common to several of the major lottery games around the world. In purchasing a ticket, players are asked to guess which of the numbers 1 through 33 will appear from a uniform distribution drawn without replacement. Representativeness suggests that players will believe that 6-number sets which are more evenly distributed across the 1 to 33 ordered line (for example, {5 9 13 18 25 30}) are more likely than 6-number sets comprised heavily of either low or high numbers (for example, {1 3 5 6 9 12}). Our hypothesis which follows directly from representativeness bias, is contradictory to the reality that both 6-number sets have equal probabilities of being drawn in the lottery.

As an "extreme" thought-experiment, consider how coincidental or unlikely the lottery outcome {1 2 3 4 5 6} may appear to players. According to representativeness, players doubt the {1 2 3 4 5 6} outcome because it does not resemble the spread of probability weight reflected in the Uniform[1,33] distribution, but rather it appears to put all the probability weight on the low number range. By representativeness, we suggest that players believe that a number set should be similar to its source distribution in terms of the density function shape, including the use of distributional moments as benchmarks. The case of {1 2 3 4 5 6} performs poorly on most measures (ex. mean, variance, skewness).

Using data on player's number choices on the national lottery game in China, which follows the exact aforementioned structure, we find that players are most likely to choose number sets which are well-spread out, in accordance with having to "represent" the uniform distribution in the drawing of each 6 number set. Lottery players over-select such evenly distributed number sets compared to their true theoretical probability, while under-selecting non-representative distributed number sets.

We examine two possible ways that players may implement representativeness. First, we consider a heuristic in which gamblers choose 2 numbers from each of the following bins [1,11], [12, 22], [23,33]. This rule of thumb, which we call the Three Bin Strategy, is promoted in instructional websites and other advice literature to lottery players. The second implementation approach we consider is a more direct behavioral optimization which we call Even Spacing Index Strategy; evenly spreading out the number picks as much as possible such that the number of unselected integers or "gaps" between each two numbers selected is approximately equal to other gaps in the set. We find that in both of these

metrics, gambler's selection of lottery number sets is more representative of the uniform distribution than implied by the distribution of these metrics in the random draws of the actual lottery game. In the context of the lottery game's pari-mutuel structure where winning tickets must share the jackpot, this bias in number set selection is at the cost of reducing their payout in the event that their number picks actually win.

Our paper contributes to the literature on belief biases in the field. Clotfelter and Cook (1993) document the Gambler's Fallacy in the Maryland "Pick 3" game, in the time dimension, finding that after a particular number appears on the winning ticket, the amount of money bet on that number falls sharply. Terrell (1994) examines a pari-mutuel lottery, the New Jersey 3-digit number game, also finding evidence for the Gambler's Fallacy, in spite of the fact that there is an expected payoff benefit to choosing "against" the fallacy. Both of these studies focus on the traditional time-dimensional aspect of Gambler's Fallacy. As in the previous two studies, we exploit the well-defined probability structure of the lottery game to demonstrate the existence of biased beliefs. Our study is the first to our knowledge, to rigorously examine the Gambler's Fallacy in the cross-sectional dimension using actual field data.[2]

We note that the traditional heads or tails example of Gambler's Fallacy is also consistent with this cross-sectional argument, in that guessers expect the total realized draws (HHH_) when considered as a set, to reflect the binomial 50-50 probability distribution, and T is required to make the realized set look more representative. Like the time-dimensional Gambler's Fallacy, the cross-sectional Gambler's Fallacy has potentially wide implications for decision-making outside the lottery domain whenever sets of items must be chosen by a decision-maker, and there is a known underlying distribution. An immediate example of this is in making forecasts about sets of outcomes. In this paper, we focus on proving the existence of this cross-sectional representativeness bias from the Gambler's Fallacy using the data on lottery number choices, but we provide some discussion of further consequences in the conclusion.

Our findings also bear resemblance to a strand of literature on the diversification heuristic, which shows that individuals have a tendency to diversify their choices in consumption choices and personal investment behavior. Simonson (1990) shows that individuals tend to diversify their product choices more when they make multiple selections simultaneously, compared to when they make choices in sequence. In testing the causes of such variety-seeking behavior, Read and Loewenstein (1995) find that time contraction and choice bracketing are the most plausible explanations. In the personal investment realm, Benartzi and Thaler (2001) find that individuals' allocation choices in defined contribution savings plans, tend to spread investments evenly over the available options, regardless of the actual features of the investment options.

Our study presents a behavior similar to the diversification heuristic: the documentation of individuals' tendency to spread or 'diversify' their choices in a set of numbers (in our case, over a number interval) more than a classical analysis predicts they would. Although our study shares the documented behavior in common with literature on the diversification heuristic, there is a distinction in the proposed mechanism for the diversifying behavior. There is a strong case (due to the simple and transparent lottery game structure) that the diversification in our study is derived from the misperception of probabilities of sets of numbers being drawn, in line with Representativeness Bias. Nevertheless, future work is needed to further distinguish and discover the links between belief biases and

[2] Haigh (2008) provides a summary and discussion of several patterns in gambler choices in lottery number picks, such as favoring certain number combinations, choosing previous winning combinations, choosing numbers which form particular patterns on the lottery ticket sheet, etc.

diversification behavior more generally.

The remainder of the paper proceeds as follows: Section 2 describes the lottery game and the data; Section 3 describes our empirical strategy; Section 4 shows the results; Section 5 concludes and suggests possible future directions for research.

2. The SSQ Lottery Game and Data

The rules of the SSQ Lottery, the national lottery game in China, are similar to those of other popular lotteries, such as the Powerball in the US and the LottoMax in Canada. The SSQ lottery is the most popular of the government operated lottery games in China. Our data was collected from Taobao Lottery, an online website where consumers can purchase SSQ lottery tickets of their choice through Taobao.com, their affiliate, the largest official online retailer of SSQ lottery tickets.

Each ticket is sold for 2 RMB. SSQ Lottery requires players to pick numbers from two groups of numbers. In the first group, called "red numbers", players need to pick 6 numbers from the range 1 to 33. In the second group, called "blue number", players need to pick 1 number from the range 1 to 16. The red numbers are drawn from the integers in Uniform[1,33] without replacement, and players' number set selections are also restricted to this criterion.[3] The winning lottery numbers are drawn randomly from the aforementioned distribution, and like the lottery games in the US, the process of choosing the numbers via machine is televised to verify authenticity.

To win the first prize jackpot, a player needs to match all 7 numbers randomly drawn as the winning number combination. The SSQ has 6 levels of prizes, depending on the number of balls matched. The details of the prize structure are shown in Table 1. The first and second prizes are pari-mutuel, depending on the number of winning tickets and the current size of the prize pool. The third to sixth prizes are fixed reward prizes, regardless of the number of winning tickets or prize pool. Each 6 number combination that the players are allowed to select, has an equal chance of satisfying any of the award prize criteria. However, for the first and second prizes, the payout of an individual player depends on how many other tickets of that specific number combination were sold; i.e. the more players who have purchased tickets with the winning number combination, the more people must share the prize.

Since only one blue number is drawn each round, from a separate distribution than the red numbers, the blue number is not suitable for testing our theory about tendencies in choosing sets of numbers in the cross-section. Therefore, our analysis focuses only on players' choice of red numbers.

Our data were gathered directly from the Taobao Lottery website over 15 rounds during the dates from Nov 11th 2011 to Dec 20th 2011. We observe the volume of tickets sold on Taobao online under each number combination during this period. The data consists of over 1.6 million lottery tickets and their corresponding number combinations purchased by over 28,000 players. This corresponds to over half a million US dollars in wager amounts over our observation period. Table 2 shows the summary statistics of the data.

[3] In other words, players may not choose the number 7 twice in their 6 number set.

Table 1: SSQ Prize Policies

Award level	Winning conditions		Prize distribution
	Number of Red balls matched (out of 6)	Blue ball matched?	
First prize	6	Yes	If the rollover money from the last jackpot is less than 100 million RMB, then the grand prize jackpot winners will split the rollover from the previous draw and the 70% from the "high prize pool". If the prize is more than 5 million RMB, each winning ticket will only be worth 5 million RMB. If the rollover money from the last jackpot is at least 100 million RMB or more, there is a two part prize package. The winners split the rollover money from the previous draw and 50% from the "high prize pool", as well as 20% from the "high prize pool". With each prize, a maximum of 5 million RMB is paid (total of 10 million RMB).
Second prize	6	No	To split the 30% of "high prize pool".
Third prize	5	Yes	Fixed amount of 3000 Yuan per winning lottery ticket
Fourth prize	5	No	Fixed amount of 200 Yuan per winning lottery ticket
	4	Yes	
Fifth prize	4	No	Fixed amount of 10 Yuan per winning lottery ticket
	3	Yes	
Sixth prize	2	Yes	Fixed amount of 5 Yuan per winning lottery ticket
	1	Yes	
	0	Yes	

Table 2: Summary Statistics

Number of Lottery Players	**28,474**
Number of Rounds	15
Total Number of Lottery Tickets Purchased	1,673,362
Total Amount of Wager (RMB)	3,346,724

	Mean	**Min**	**Max**	**SD**
# of tickets per round	112,108	94,595	135,396	10,597
# of tickets per lottery player (all 15 rounds)	59	4	38,084	397
# of tickets per lottery player per round	14.7	4	11,832	92

Note: The dates of the lottery game are from Nov 15th 2011 to Dec 20th 2011.

3. Empirical Strategy

We consider two approaches to gamblers' manifestation of Representativeness Bias in the cross-section. Our first approach (Three Bin Strategy) is based on common advice given to lottery players

about how to choose numbers, partitioning the number range into three sections as shown in Appendix Figure A1. We test to see whether gamblers' choice of numbers deviates from the actual likelihood of those number patterns being chosen. Our second approach, the Even Spacing Index, tests the approximation of the uniform distribution more directly, by examining the degree to which players are attracted to number combinations which are more 'optimally' or evenly spread out over the range 1 to 33.

The Three Bins approach can be considered as possible rule of thumb or heuristic for Even Spacing behavior. We test the Three Bins approach first, since we are already aware that this number selection strategy is promoted in some gambling advice literature. To test whether players have more *generally* cross-sectional representative beliefs aside from this heuristic, we then check the degree to which number choices are spread out evenly across the number range using the Even Spacing Index.

3.1 Three Bins Strategy

Our inspiration for the Three Bins test comes from advice frequently dispensed to lottery players online and in other advice literature. According to this approach, players are advised to choose 2 numbers from each of three bins in the relevant number range. An example of such analysis found online is shown in the Appendix in Figure A1.

We divide the integers 1 to 33 into three bins: [1 - 11], [12 - 22], [23 - 33]. We use the following notation to denote a lottery number combination for a single $Ticket_i$:

$$Ticket_i = \{ r_1, r_2, r_3, r_4, r_5, r_6 \}$$

Let E_1 represent the amount of chosen numbers which fall into the bin of [1 – 11]; Let E_2 represent the amount of chosen numbers which fall into the bin of [12 – 22]; Let E_3 represent the amount of chosen numbers which fall into the bin of [23 – 33]. The vector of (E_1 , E_2 , E_3) is a measure for how evenly the numbers are picked across the three bins.

For example, for $Ticket = \{1, 2, 3, 4, 5, 6\}$, the vector is (6,0,0) since all of the numbers are falling into the bin of [1–11]. On the other hand, for $Ticket = \{3, 9, 12, 17, 26, 30\}$, the vector will be (2,2,2), since this set of numbers contains two numbers from each bin.

Players selecting numbers in this way may use the bins as a rule of thumb for representativeness. To choose a set of numbers which has even spacing all around, still takes some effort and thinking, whereas the Three Bins approach is conceivably more automatic. Note that an alternative way to divide the bins might be by intervals of 10 instead of 11, however the results would be quite similar under either assumption.[4]

To test the Three Bins strategy, we compare the empirical frequency of these different possible bin vectors in gamblers' actual number choices, to the theoretical frequencies as determined by the lottery's uniform random draw. The results are given in Section 4.

3.2 Even Spacing Index Approach

In this subsection, we create another index to measure the even distribution of numbers choices in lottery tickets. We create an Spacing Index, where we use the sum of the squares of number gaps between the six chosen numbers as a measurement of spacing disparity. The more evenly the numbers are distributed, the smaller this sum of the squares will be.

[4] Under this alternative arrangement 31, 32, and 33 would be combined with the bin [21,30]. Our current arrangement allows for equally sized bins, and is also the bin definition used in most of the advice literature (see Appendix Figure A1).

More specifically, suppose we use the following notation to denote a lottery number combination with numbers listed in sequential order, for a single $Ticket_i$:

$$Ticket_i = \{ r_1, r_2, r_3, r_4, r_5, r_6 \}$$

The spacing index is defined as follows, calculating the sum of squares of the number of integers between the numbers chosen:

$$Spacing\_Index_i = (r_2 - r_1 - 1)^2 + (r_3 - r_2 - 1)^2 + (r_4 - r_3 - 1)^2 + (r_5 - r_4 - 1)^2 + (r_6 - r_5 - 1)^2 + (r_1 - 1 + 33 - r_6)^2$$

For instance, the non-representative ticket {1, 2, 3, 4, 5, 6} will have a Spacing Index of 729 using the above formula while the smallest Spacing Index is 123. One number combination out of several which achieves this smallest index value, is {2, 7, 12, 18, 24, 30}. The Spacing Index is a convenient way to summarize the cross-sectional representativeness characteristics of different number sets. For example, all possible choices of *consecutive* numbers in a lottery ticket would be categorized as extremely unbalanced and would have the same Spacing Index value of 729.

Compared to the Three Bins approach, the Spacing Index captures further detail about players' choices compared to the Three Bins approach, by informing us whether players prefer to spread their lottery number choices within bins as well as across bins, and whether in general, more evenly spaced number sets are more attractive, beyond the confines of the bins.

4. Results

4.1 Three Bins Results

Our main results for the Three Bin approach can be seen by looking at Figures 1 and 2. Figure 1 shows the theoretical probabilities of choosing particular 3 bin number combinations given a uniform draw of 1 to 33. Bin types are displayed on the horizontal axis, in order from more representative on the far left (2,2,2) moving towards (weakly) less representative bin types towards the right hand side. Figure 2 shows the empirical deviation from these theoretical proportions in percentage points.

Although theoretically, a 222 combination is indeed more likely to be chosen than the other combinations, we can see from Figure 2 that 222 is *disproportionately* favored in the data. That is, while vector 222 should come up about 15% of the time in theory, it is in fact chosen by gamblers over 19% of the time. This difference is statistically and economically significant. Gamblers' overall likelihood of choosing "unbalanced" or uneven distribution combinations of numbers is less than the theoretical probability for less balanced 3-bin vectors such as (0,2,4), (0,3,3), (0,4,2), (1,1,4), (4,1,1) and others, while being higher than the theoretical probability for more balanced vectors such as (2,2,2), (1,3,2), (2,1,3), (2,3,1), and (3,1,2).

Based on Figure 2, we can additionally see that there is an approximate turning point – players overly prefer bin allocations containing 3, 1, and 2, while overly-avoiding any combination less evenly distributed than this. Here we do not attempt to accommodate all the behavioral biases in the number picking behaviors, but focus on the general trend in over-selection and under-selection as a function of set representativeness. There are a few exceptions to the overall trend (for example bin (1,2,3) and bins (0,0,6) and (6,0,0)). However, the pattern holds that gamblers generally over-gravitate towards more representative number sets, and under-gravitate towards less representative sets.

We conduct formal statistical tests to confirm the claims above. The complete results,

disaggregated by each of the 15 lottery rounds we observe, are displayed in Table 2. We first implement the Chi-Square goodness of fit test to examine whether the observed relative frequencies of the number combination choices differ from the theoretical distribution. The χ^2 value is over 1000, well beyond the critical value of 47 (for 27 degrees of freedom) at the 99% confidence level, and easily rejects the null hypothesis that the actual frequency of 3-bin number combinations is the same as the theoretical probability under random number choice, shown in Table 2.[5]

However, important information about the actual differences between the theoretical probability distribution and the actual one, will be missed if only the Chi-Square Goodness of Fit Test is employed on the entire distribution of number picks. We further test for the difference between each theoretical probability and the mean of the actual fifteen-round frequencies of occurrence on *each* 3-bin number combination with a simple t-test.

To be specific, we use p_b to denote the theoretical probability for a specific 3-bin number combination indexed by b. Altogether, there are 28 3-bin number combinations, and we conduct a separate t-test for each. We let $f_{b,r}$ represent the actual frequency of the 3-bin combination b in round r. In our data, $r \in \{135, 136, 137, \ldots, 148, 149\}$. We want to test, given all other factors fixed, for the 3-bin number combination b, whether the sample mean $f_{b,r}$ for these fifteen rounds of lottery games is close to the theoretical probability, p_b. The results are shown in Table 2. Overall, the bin level results confirm the results that we find in the aggregate. The significance and direction of the test results on the 3-bin number combinations reinforce the claim shown in Figure 2, that players overly gravitate to bin allocations that are more evenly distributed while overly avoiding combinations that are less evenly distributed.

While it is clear from these results that people disproportionately tend to pick numbers which are equally allotted among the three bins (222), or nearly as representative bin-combinations, a question is whether players have a more *general* tendency to want to spread out their number choices evenly across the range of possible numbers, beyond the 3-bin heuristic. To address this question, we turn to the Spacing Index results.

[5] There are 27 degrees of freedom given the 28 different 3-bin categories.

Figure 1: Theoretical Bin Probabilities

y-axis: percentage points, x-axis - Three Bin vectors

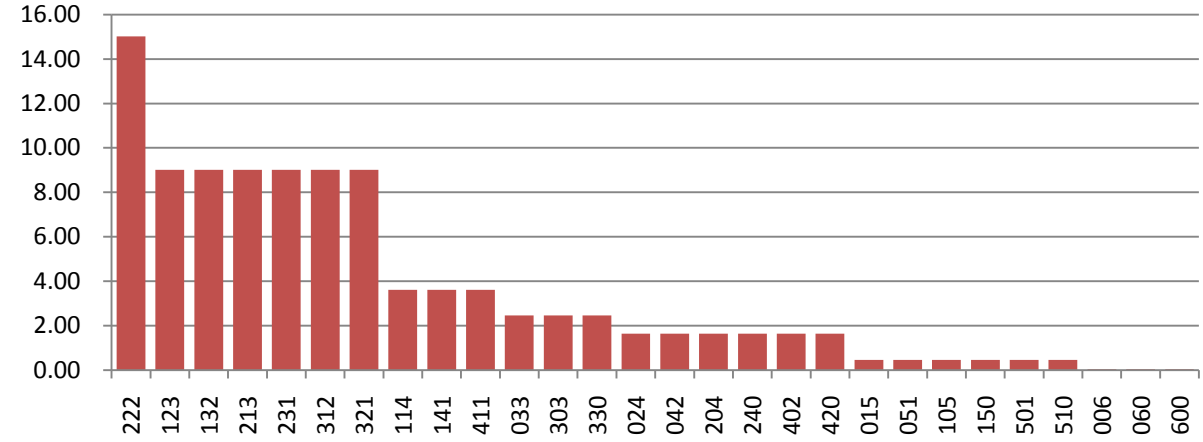


Figure 2: Empirical Over-Selection of Bins

y-axis: percentage points, x-axis - Three Bin vectors

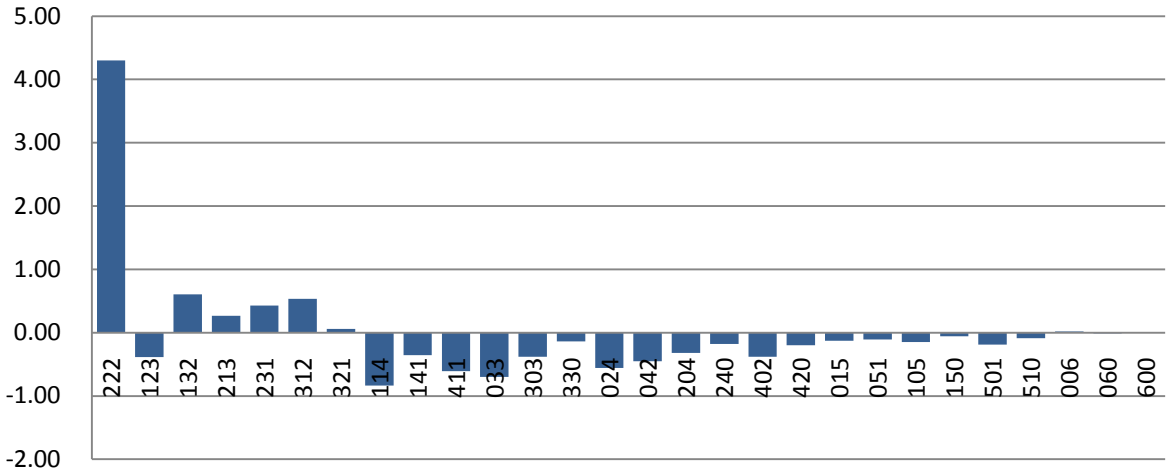


Table 2: Frequency of Number Picking Strategies

3-bin combination	Theoretical Probability %	Actual Distribution																	
		Aggregate (All rounds)			Each Round (%)														
		%	Diff	# of Tickets	#135	#136	#137	#138	#139	#140	#141	#142	#143	#144	#145	#146	#147	#148	#149
006	**0.04**	**0.06**	**0.02***	**1,068**	0.05	0.07	0.07	0.12	0.05	0.08	0.06	0.06	0.06	0.06	0.05	0.05	0.07	0.05	0.05
015	**0.46**	**0.33**	**-0.13***	**5,502**	0.26	0.29	0.31	0.40	0.27	0.28	0.28	0.37	0.32	0.29	0.34	0.50	0.40	0.34	0.26
024	**1.64**	**1.08**	**-0.56***	**18,164**	0.91	1.08	1.08	1.00	1.07	1.29	1.20	1.34	0.87	0.99	1.27	1.04	0.95	1.14	0.86
033	**2.46**	**1.76**	**-0.70***	**29,516**	1.98	1.74	1.71	1.91	1.70	1.61	1.59	1.70	1.56	1.88	2.19	1.91	1.75	1.66	1.47
042	**1.64**	**1.19**	**-0.45***	**20,011**	1.07	1.37	1.11	1.24	1.54	1.07	1.03	1.21	0.95	1.05	1.44	1.15	1.01	1.39	1.13
051	**0.46**	**0.35**	**-0.11***	**5,892**	0.28	0.52	0.31	0.28	0.57	0.33	0.32	0.31	0.28	0.48	0.35	0.34	0.28	0.29	0.25
060	**0.04**	**0.03**	**-0.01***	**519**	0.04	0.05	0.03	0.04	0.03	0.01	0.03	0.03	0.03	0.02	0.03	0.02	0.03	0.06	0.03
105	**0.46**	**0.31**	**-0.15***	**5,214**	0.33	0.36	0.36	0.32	0.34	0.29	0.27	0.31	0.30	0.28	0.27	0.37	0.29	0.29	0.27
114	**3.6**	**2.77**	**-0.84***	**46,643**	3.08	2.54	2.60	2.67	2.63	2.50	2.64	2.78	2.44	2.72	3.32	2.95	2.88	3.06	2.80
123	**9.01**	**8.63**	**-0.38***	**145,055**	8.66	7.62	8.37	8.47	7.82	8.45	8.28	8.97	7.80	8.41	9.92	9.60	9.01	9.52	8.55
132	**9.01**	**9.62**	**0.61***	**161,706**	9.70	9.45	9.50	10.12	11.08	9.23	9.56	9.55	8.68	9.03	10.08	9.88	10.02	9.58	8.68
141	**3.6**	**3.25**	**-0.36***	**54,580**	3.21	3.42	3.12	3.39	3.18	3.30	3.45	3.37	3.15	3.20	3.21	2.93	3.13	3.17	3.37
150	**0.46**	**0.40**	**-0.06***	**6,776**	0.37	0.55	0.26	0.42	0.45	0.36	0.38	0.33	0.38	0.43	0.40	0.47	0.51	0.38	0.35
204	**1.64**	**1.32**	**-0.32***	**22,202**	1.49	1.38	1.32	1.23	1.46	1.38	1.30	1.58	1.28	1.19	1.24	1.19	1.01	1.45	1.23
213	**9.01**	**9.28**	**0.27**	**156,110**	10.32	8.94	9.15	9.29	8.65	9.72	9.29	8.99	9.01	9.43	9.38	10.06	9.25	9.21	8.89
222	**15.02**	**19.32**	**4.30***	**324,910**	18.99	18.10	19.68	19.87	22.38	19.18	19.33	20.58	19.02	19.28	18.66	18.37	18.85	18.39	18.74
231	**9.01**	**9.44**	**0.43***	**158,667**	9.46	9.32	9.76	9.39	8.66	9.92	10.07	8.98	10.17	8.63	8.87	8.65	10.46	9.38	10.12
240	**1.64**	**1.46**	**-0.18***	**24,606**	1.31	1.95	1.38	1.35	1.39	1.29	1.47	1.28	1.44	1.73	1.49	1.44	1.67	1.21	1.51
303	**2.46**	**2.08**	**-0.38***	**35,006**	2.32	2.27	1.93	2.00	2.07	2.06	1.91	1.77	2.07	2.37	2.21	2.04	2.03	2.06	2.17
312	**9.01**	**9.55**	**0.54***	**160,537**	9.53	9.81	9.80	9.60	8.17	10.01	9.34	9.88	11.25	9.95	8.52	9.03	8.87	9.56	9.95
321	**9.01**	**9.07**	**0.06**	**152,483**	8.53	9.28	9.12	8.98	8.54	9.47	9.43	8.72	9.85	9.32	8.16	9.07	8.66	9.28	9.70
330	**2.46**	**2.32**	**-0.14***	**38,965**	2.05	2.75	2.39	2.14	2.02	1.94	2.60	2.14	2.33	2.35	2.40	2.45	2.42	2.22	2.53
402	**1.64**	**1.26**	**-0.38***	**21,115**	1.30	1.46	1.14	1.28	1.19	1.29	1.23	1.14	1.34	1.37	1.31	1.18	1.25	1.27	1.07
411	**3.6**	**3.00**	**-0.61***	**50,508**	2.81	3.39	2.93	2.73	2.85	2.80	3.02	2.58	3.28	3.27	2.79	2.91	3.10	2.99	3.64
420	**1.64**	**1.44**	**-0.20***	**24,228**	1.28	1.53	1.69	1.20	1.28	1.30	1.32	1.26	1.52	1.60	1.42	1.77	1.52	1.44	1.53
501	**0.46**	**0.27**	**-0.19***	**4,612**	0.27	0.36	0.26	0.24	0.26	0.44	0.24	0.26	0.22	0.29	0.25	0.22	0.21	0.26	0.32
510	**0.46**	**0.37**	**-0.09***	**6,208**	0.36	0.39	0.54	0.32	0.30	0.29	0.33	0.42	0.38	0.35	0.40	0.38	0.30	0.33	0.48
600	**0.04**	**0.05**	**0.01**	**817**	0.05	0.04	0.09	0.03	0.03	0.10	0.03	0.09	0.03	0.04	0.04	0.04	0.05	0.03	0.05
Total				**1,673,362**															

* *$p < 0.05$, rejects the null hypothesis that the sample mean over the 15 rounds, for 3-bin number combination is the same as the theoretical probability.*

4.2 Even Spacing Index Results

As in the Three Bins approach, we first need to consider the theoretical probabilities of different Spacing Index values arising under the assumption of random number selection from a uniform distribution. Figure 3 shows the theoretical distribution of Spacing Index values under the assumption of randomized number selection. The shape of this distribution is determined by the natural frequency of Spacing Index values arising from all the possible combinations of 6 number lottery tickets, where each 6 number combination has an equally likely chance of being chosen.

Figure 4 shows the difference between the *observed empirical* probability and the *theoretical* probability. The results are clear and similar to the Three Bin results - the difference between the empirical and theoretical distribution is systematic. The empirical probabilities skew towards smaller Spacing Index values. In other words, people are more likely to choose very evenly distributed numbers, and less likely to choose unevenly distributed ones compared to what is predicted by the actual lottery drawing process.

Figure 3: Theoretical Probabilities of the Spacing Index Values

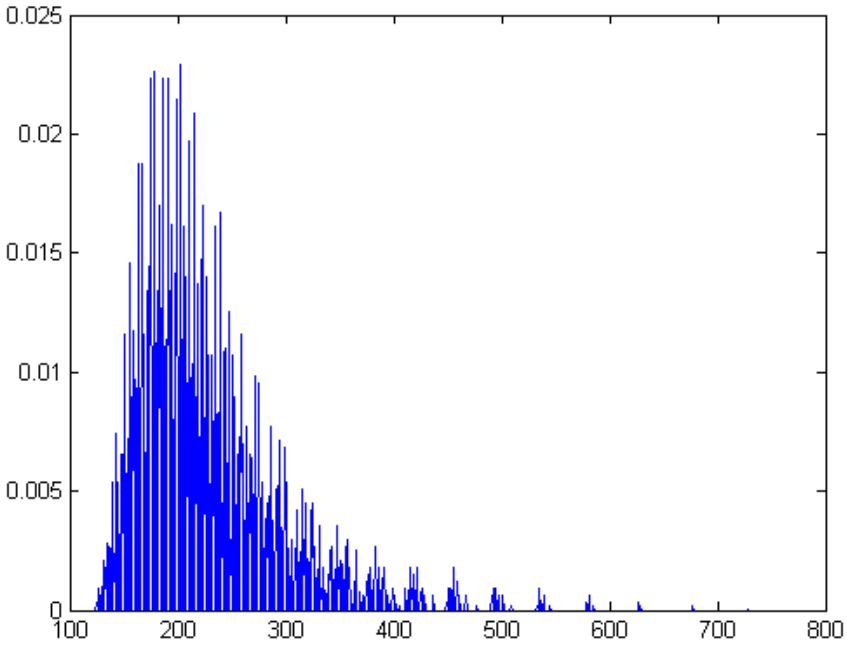
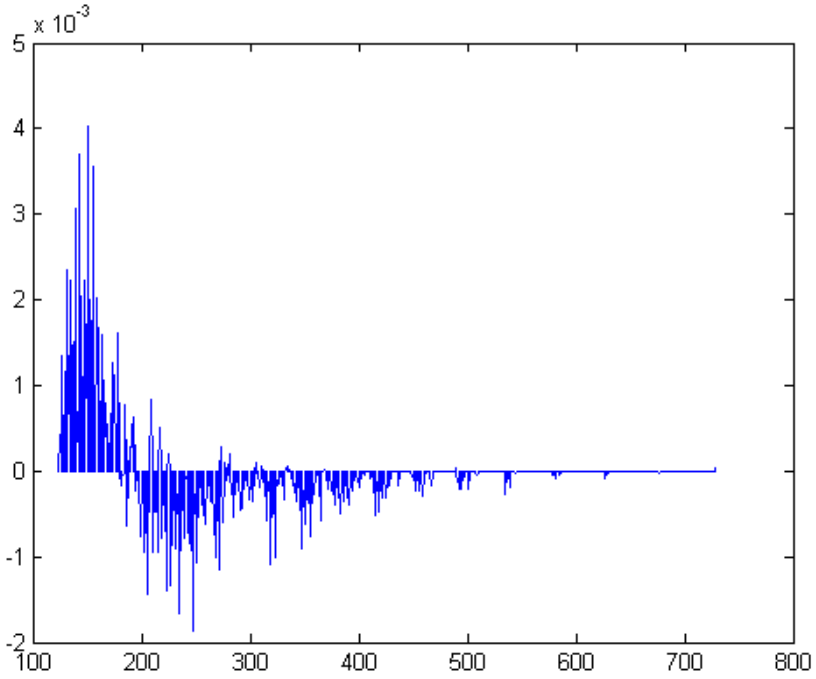


Figure 4: Empirical Over-selection of Spacing Index Values



Similar to our findings in the Three Bins approach, there appears to be a turning point, around the Spacing Index value of 200 where gamblers begin disfavoring less representative number combinations. An example of a number set which has a Spacing Index value of 200 is {2,4,14,15,26,30}. As the Spacing Index increases, people almost never revert to over-selecting those unevenly spaced number combinations. This pattern is remarkably monotonic, especially when one considers possibilities such as players' tendencies to pick favorite or lucky numbers. Figure 4 suggests that if players do choose such numbers in their lottery ticket, they are likely to choose the other numbers in the ticket in a way which accommodates even spacing.[6]

The degree of over/under-selection can be understood in terms of the probability magnitudes. The theoretical probability of each Spacing Index value is in the range of 0.01 to 0.02, while the magnitude of the absolute difference between the empirical and theoretical probabilities is in the range of 0.001 to 0.005. We also implement the Chi-Square goodness of fit test here, to test the null hypothesis that the theoretical probability distribution of the Spacing Index is the same as the empirical frequency distribution of the Spacing Index. The test easily rejects the null hypothesis with a χ^2 value of over 1,000, which far exceeds the critical value of 218 (for 172 degrees of freedom), at the 99% confidence level.[7]

5. Conclusion

In this paper we have provided clear evidence of a particular belief fallacy, a "cross-sectional Gambler's Fallacy" in decisions about how to choose sets of numbers. Our data are from one of the

[6] The results suggest for example, the in the case where a player wants to choose his favorite number 9, he may use 9 as the anchor around which to choose the other numbers in his ticket based on even spacing. In the 3-bins approach, he may allow himself to choose just one or two other numbers in the [1,11] bin.

[7] Here, there are 172 degrees of freedom due to 173 different values of the Spacing Index.

largest and most popular lottery games in the world, and consist in our observation period, of over 1.6 million tickets purchased and over 28,000 individuals. Gamblers gravitate more heavily towards cross-sectionally representative number sets than the actual probabilities suggest, and they disfavor picking number sets which appear unrepresentative. We are able to cleanly detect this fallacy due to the transparent probability structure of the lottery game.

Furthermore, the lottery we examine is pari-mutuel. This means that belief fallacies in the game are at an expected cost to gamblers. Conditional on any given lottery number combination winning, under the cross-sectional Gambler's Fallacy, a lottery player is more likely to have chosen a number combination which is popular among many other players. The implication is that more people will be splitting the jackpot. On the other hand, consider the case where a player picks a non-representative number combination which is relatively unpopular among other gamblers. If their combination wins, they will receive a larger payout or fraction of the jackpot. As Terrell (1994) points out, this makes such representativeness-driven belief fallacies about more than just marginal preferences for certain numbers, but shows that these beliefs can be monetarily disadvantageous.

The belief bias that the cross-sectional Gambler's Fallacy represents is likely to have consequences for other settings besides lotteries. For example, consider the case of a store manager who needs to estimate the distribution of customers that will visit a store on a random day, based on their ice cream flavor preference (ex. strawberry, vanilla, chocolate). Suppose that the manager is told by the owner about the approximate distribution of flavor preferences among customers, which serves as the manager's 'theoretical' reference (for example: 1/3 strawberry, 1/3 vanilla, 1/3 chocolate). Cross-sectional representativeness as found in this study, would suggest that the manager will tend to hold *overly strong* beliefs that the clients arriving on any particular day should closely reflect this distributional information the owner told him. On the other hand, he would tend to find it very doubtful (more doubtful than implied by actual probabilities) that 40 out of 50 customers arriving in a single day will prefer strawberry ice cream.

Finally, we would like to note that the ability of cross-sectional representativeness to predict choice behavior could potentially depend on the features of the underlying distribution. Due to our field data being from a lottery setting, our analysis is limited to detecting set representativeness in the case where the source distribution is uniform. Our results also bear a close resemblance to the literature on the diversification heuristic in this setting, in that we find a tendency of decision-makers to 'overly spread out' their choices. More work is needed to establish the potential links between diversification behavior and our findings here, as well as the robustness of set representativeness to different theoretical distributions. We leave the investigation of these issues to our future work.

References:

Benartzi, Shlomo and Richard H. Thaler, "Naïve Diversification Strategies in Defined Contribution Saving Plans", American Economic Review, Vol. 91 (2001), No. 1, p. 79 – 98.

Clotfelter, Charles T., and Philip J. Cook, "The 'Gambler's Fallacy' in Lottery Play", Management Science, Vol. 39, No. 12 (Dec., 1993), p. 1521 – 1525.

Croson, Rachel and James Sundali, "The Gambler's Fallacy and the Hot Hand: Empirical Data from Casinos," The Journal of Risk and Uncertainty, 30:3, p. 195 – 209, 2005.

Galbo-Jorgensen, Claus B., Sigrid Suetens, and Jean-Robert Tyran, "Predicting Lotto Numbers", Working Paper, August 2012.

Grote, Kent R., and Victor A. Matheson, "The Economics of Lotteries: A Survey of the Literature", College of the Holy Cross Working Paper No. 11-09, August 2011.

Haigh, John, "The Statistics of Lotteries", Handbook of Sports and Lottery Markets (Chapter 23), Elsevier B.V., 2008.

Rabin, Matthew, "Inference by Believers in the Law of Small Numbers," Quarterly Journal of Economics, August 2002.

Rabin, Matthew and Dimitri Vayanos, "The Gambler's and Hot-Hand Fallacies: Theory and Applications," The Review of Economic Studies (2010), Vol.77, p. 730 – 778.

Read, Daniel and George Loewenstein, "Diversification Bias: Explaining the Discrepancy in Variety Seeking Between Combined and Separated Choices", Journal of Experimental Psychology: Applied, Vol.1 (1995), No. 1, p. 34 – 49.

Simonson, Itamar, "The Effect of Purchase Quantity and Timing on Variety-Seeking Behavior", Journal of Marketing Research, Vol. 27 (1990), No. 2, p. 150 – 162.

(This paper is not cited in our manuscript. Should be deleted?)

Terrell, Dek, "A Test of the Gambler's Fallacy: Evidence from Pari-mutual Games", Journal of Risk and Uncertainty, Vol. 8, 1994, p. 309 – 317.

Tversky, Amos, and Daniel Kahneman, "Belief in the Law of Small Numbers", Psychological Bulletin, 1971, Vol. 76, No. 2, p. 105 – 110.

Appendix:

Figure A1: (Example) Keeping track of lottery outcomes using the 3-bin approach

The following figure from a website (http://trend.baidu.lecai.com/ssq/redThreeAreaTrend.action?onlyBody=false) illustrates the 3-bin approach to lottery number selection. While this website tracks the trends in lottery numbers cumulatively over rounds, it encourages players to think of the number frequencies in terms of the 3 bins (the wide columns marked with Chinese characters一区,二区,三区) and even allocation of numbers across bins.

期号 ↑	一区											二区											三区										
	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
2013095	01	1	17	1	5	06	3	2	7	14	3	6	8	1	15	9	16	1	19	1	3	1	5	4	4	2	17	28	29	13	2	10	20
2013096	01	02	18	2	6	1	4	3	8	15	4	7	9	2	1	10	17	2	1	2	4	22	6	5	5	3	18	28	29	30	3	11	21
2013097	1	1	19	3	05	2	5	4	9	16	5	8	10	14	2	11	17	3	2	3	5	22	23	6	25	4	19	1	1	1	4	12	22
2013098	2	2	20	4	1	3	07	5	10	17	6	9	11	1	15	12	1	18	19	20	6	1	1	7	1	26	20	2	2	2	5	13	23
2013099	3	3	21	5	05	4	1	6	11	18	11	10	12	2	1	13	2	1	1	20	21	2	2	8	2	26	21	3	3	3	31	14	24
2013100	4	4	22	04	1	5	2	08	12	19	11	11	13	14	2	16	3	2	2	20	1	3	3	9	3	1	22	4	4	4	1	15	25
2013101	5	5	23	1	05	6	07	1	09	20	1	12	14	1	3	1	4	3	3	1	2	4	23	10	4	2	27	5	5	5	2	32	26
2013102	6	02	24	04	05	06	1	08	1	21	2	13	15	2	4	16	5	4	4	2	3	5	1	11	5	3	1	6	6	6	3	1	27
2013103	7	02	25	04	1	1	2	1	09	22	3	14	13	3	5	1	6	18	5	20	4	6	2	12	6	4	2	7	7	7	4	2	28
2013104	01	02	26	04	2	2	3	2	1	23	4	15	1	4	15	2	17	1	6	1	5	7	3	13	7	5	3	28	8	8	5	3	29
2013105	01	1	27	1	3	3	4	3	2	24	11	16	2	5	1	3	1	2	7	2	6	8	23	14	8	6	27	1	9	9	31	32	30
2013106	1	2	28	2	4	4	5	4	09	25	11	17	3	6	2	4	2	3	8	3	7	9	23	15	9	7	1	2	10	30	31	32	31
2013107	2	3	29	3	5	5	07	5	09	26	11	18	4	7	3	5	17	4	9	4	8	10	1	16	10	8	2	28	11	1	31	1	32
2013108	3	4	30	4	6	6	1	6	1	27	1	19	5	8	4	16	1	5	10	5	21	22	2	17	11	9	3	28	12	2	31	32	33
2013109	4	5	31	5	7	7	2	7	09	28	2	20	6	9	5	1	2	6	11	6	1	1	23	24	12	10	27	1	29	3	1	32	34
2013110	5	6	32	6	8	8	3	8	1	29	3	21	7	10	15	2	17	18	12	7	21	2	1	1	13	11	1	2	29	4	2	32	35
2013111	01	02	03	7	9	06	4	08	2	30	4	22	8	11	1	3	1	1	13	8	1	3	2	2	14	12	2	3	1	5	3	1	33
2013112	01	1	1	8	10	06	5	1	3	31	5	12	13	12	2	4	2	2	14	9	2	22	3	3	15	13	3	4	2	6	31	2	1
2013113	1	2	2	04	11	1	07	2	4	32	11	1	1	13	3	5	17	3	15	10	3	1	4	24	16	14	4	5	3	7	1	3	33
2013114	2	3	3	04	12	06	1	3	5	33	1	2	2	14	4	6	17	4	16	11	21	2	23	1	17	15	5	6	4	8	2	4	33
2013115	3	4	03	1	13	1	2	4	6	34	2	12	3	15	5	16	17	18	17	12	1	3	1	2	18	16	27	7	5	9	3	5	1
2013116	4	5	1	2	14	2	3	5	7	35	3	12	4	16	15	1	1	1	18	13	21	4	2	3	19	26	1	8	6	10	4	32	33
2013117	5	6	2	3	15	3	4	6	09	36	4	12	13	17	1	2	2	2	19	14	1	5	3	24	20	1	27	9	7	11	5	1	33
2013118	6	02	03	4	16	4	5	7	1	37	5	1	1	18	2	3	17	3	20	15	2	22	4	1	21	2	1	10	8	12	6	32	33