How Lies Induced Cooperation in "Golden Balls:" A Game-Theoretic Analysis

Brams, Steven J. and Mor, Ben D.

New York University, University of Haifa

December 2019

How Lies Induced Cooperation in "Golden Balls:" A Game-Theoretic Analysis

Steven J. Brams
Department of Politics
New York University
New York, NY 10012
USA
steven.brams@nyu.edu

Ben D. Mor
School of Political Sciences
Division of International Relations
University of Haifa
Haifa, Israel
b.mor@poli.haifa.ac.il

Abstract

We analyze a particular episode of a popular British TV game show, “Golden Balls,” in which one of the two contestants lied about what he intended to do, which had the salutary effect of inducing both contestants to cooperate in what is normally a Prisoners' Dilemma (PD), wherein one or both contestants usually defected. This “solution” to PD assumes that the liar desired to be honorable in fulfilling his pledge to split the jackpot if he won but, surprisingly, he achieved this end without having to do so, astonishing the audience and receiving its acclaim. We note that this action has a biblical precedent in King Solomon’s decision to cut a baby in two.

Introduction

Beginning in June 2007 and running for more than two years, a popular British daytime game show called "Golden Balls" led to substantial frustration for many contestants, who were lied to and often betrayed by the other contestants.[1] This is not surprising, because in the final round of Golden Balls two contestants play a game called "Split or Steal," which, we argue, is a Prisoners' Dilemma.

What *is* surprising is that in one episode,

https://www.youtube.com/watch?v=S0qjK3TWZE8

one of the contestants announced in advance how he would play and then reneged on both his promises (in fact, we never learn whether he would have kept his second promise). Paradoxically, these lies led both players to cooperate in the original game.

In this note, we show that the player who made the announcement, whom we'll call *A*, complicated the game by pledging that he would choose a particular option in the original game and, if feasible, then make a later choice outside the original game. If the payoff to this player depends only on money and status—how did I do in comparison to my opponent?—we show in the next section that (i) the original game is a 2×2 Prisoners' Dilemma, in which cooperation is strongly dominated; and (ii) *A*'s announcement induces a 3×2 game, in which cooperation is weakly dominated and thus of no help in fostering cooperation. But if *A*'s payoffs also depend on his honor in fulfilling his second promise, the game is transformed into another 3×2 game with a completely different outcome.

---

[1] Van den Assem, van Dolder, and Thaler (2012) analyzed 287 episodes, finding that in 69% either one (44%) or both (25%) players chose Steal, leaving only 31% in which there was no defection from Split so the jackpot was split.

How did this happen if the players' strategies do not change in the second 3×2 game? First, *A*'s announcement signaled that his goal might not be only to maximize his monetary payoff. Instead, it suggested that he might wish to honor his pledge about what he said he would do (split his winnings) after play of the original game. If he does keep his promise and thereby demonstrates his "honorability," then cooperation is no longer weakly dominated but weakly dominant instead.

We begin our analysis by describing the rules of Golden Balls and then what usually happens in play of the game.[2] That *A*'s announcement in the aforementioned episode was only cheap talk is supported by the fact that *A* kept neither of his promises. Nevertheless, we show that *A*'s announcement and his failure to keep one of his promises was rational. Indeed, it laid the groundwork for cooperation in the game.

Rules of Golden Balls

After a series of preliminary rounds, four contestants are reduced to two. We focus on this final stage of the game (called "Split or Steal"), which is governed by the following rules:

- Each contestant is given two balls. When opened, one indicates Split and the other Steal.
- Each contestant secretly opens his or her ball—to determine which is Split and which is Steal—and chooses one. Before making a choice, however, the contestants may speak

[2] The TV show has attracted some scholarly attention, because in the experimental study of PD, it is thought to simulate "real life" situations more validly than laboratory settings. See, for example, Hart (2010), van den Assem, van Dolder, and Thaler (2012), Burton-Chellew and West (2012), and Turmunkh, van den Assem, and van Dolder (2019).

to each other and also ask for advice from the host.

- If both choose Split, they each receive half the jackpot.
- If both choose Steal, neither contestant wins anything.
- If one contestant chooses Steal and the other Split, the Steal contestant wins the entire jackpot and the Split contestant nothing.

We assume that the best outcome for *A* and the other contestant, whom we call *B*, is to win everything (payoff of 4), next best to win half (payoff of 3), next worst to win nothing when the opponent also wins nothing (payoff of 2), and worst to win nothing when the opponent wins everything (payoff of 1). We rank the last outcome worst because of the anger, humiliation, or shame a player would feel if he or she were betrayed into thinking an opponent would Split—but chose instead Steal—when he or she Split.[3]

These payoffs indicate only an ordinal ranking of outcomes from best to worst. As in Prisoners' Dilemma, the noncooperative strategy of Steal for each player strongly dominates the cooperative strategy of Split, rendering (2,2) the unique Nash-equilibrium outcome (starred in Game 1), at which both players obtain nothing.

Game 1

			B	
		Split		Steal
	Split	(3,3)		(1,4)
A				
	Steal	(4,1)		(2,2)*

[3] Later in the paper, we provide evidence for this preference in a post-game quote from B. In extant studies of "Split or Steal," where this assumption is not made, the game is referred to as a variant of PD or "weak" PD, because if only monetary payoffs underlie players' preferences, then being betrayed when choosing Split is as bad as mutual defection. In this case, unlike PD, the Steal strategy is weakly dominant, and there are three Nash equilibria.

In the actual play of Golden Balls on TV, each contestant usually tries to persuade his or her opponent to Split, promising to reciprocate so that both obtain half the jackpot. But this strategy is not convincing, often leading one or both players to Steal and forgoing the cooperative (3,3) outcome.

A's Announcement

In the aforementioned episode of the game, which was between two men, *A* emphatically said that that he would choose Steal. *B* chose Split, presumably because *A* had said that upon completion of the original game, he would give *B* half the jackpot if he had won, keeping the other half for himself.

This, of course, is cheap talk, because *B* has no assurance that *A* will keep his promise and give him half the jackpot. Because it is possible that *A* will do so, however, we give him a choice of keeping or not keeping his word after he Steals (as he said he would do) and *B* Splits, which yields the 3×2 game shown in Game 2.[4]

Game 2

			B	
		Split		Steal
	Split	(3,3)		(1,4)
A	Steal, then Split (if possible)	(3,3)		(2,2)
	Steal, then don't Split (if possible)	(4,1)		(2,2)*

[4] "If possible" becomes possible when *A* Steals and *B* Splits in the original game, in which case *A* can then either keep his promise and Steal or Split instead. We know of no other game-theoretic treatments that recognize that *A*, after his announcement, has three, not two, strategies. For other game-theoretic models of this specific episode of Golden Balls, see Nikolaev (2014), Talwalkar (2012), and Cornell University Course Blog (2012).

This game could also be written in extensive form (i.e., as a game tree), whereby *A* makes a decision after learning the outcome of the 2 × 2 game. But the normal form (i.e., as a payoff matrix) makes it easier to compare with the earlier 2 × 2 game and the final game we discuss in the next section. Notice that, as in Game 1, each player does best when he wins the entire jackpot (4), next best (3) when there is a split, next worst (2) when both players Steal and there is nothing to split, and worst (1) when one wins the entire jackpot and the other nothing.

Observe that *A* has a weakly dominant strategy of "Steal, then don't Split (if possible)," whereas both of *B*'s strategies are undominated. But given *A*'s weakly dominant strategy, (2,2) in the lower right is the unique Nash-equilibrium outcome (starred) in Game 2, echoing the Nash-equilibrium outcome in the 2 × 2 game in which both players Steal and, consequently, walk away empty-handed.

These strategies, however, were not the choices of the players in the TV game—quite the opposite: both Split—suggesting that the payoffs in Game 2 are not an accurate reflection of the players' preferences. Instead, we believe, *A* had another goal in mind besides maximizing his winnings.[5]

The Players' Preferences

We suggest that the effect of *A*'s announcement was not only to increase *A*'s strategies in Game 1 from two to three in Game 2. *A* also wanted to alter *B*'s perception

[5] Thereby we do not prescribe what the players should do but work backwards from their actual choices to infer what their goals must have been to act in the way that they did. In effect, we reconstruct a game in order to try to offer a coherent explanation, through "revealed preference," of why the players' choices are consistent with their actions (i.e., are rational). While this reasoning may appear tautologous, it is the foundation of all science, including mathematics, in which nonobvious theorems are derived from assumptions. Here the rational choices of *A* and *B* in a game provide an explanation of their behavior that, on first blush, seems inexplicable.

of the game first by appearing to be sincere in declaring his intention to Steal in the original game, then also saying that if *B* Split—so A would win the entire jackpot—*A* would honor his pledge to split the jackpot later.

But being honorable for *A*, we think, does not simply mean that he privately takes pride in "doing the right thing" by keeping his pledge to Split. He also wants to demonstrates publicly—in front of the studio audience as well as 2 million TV viewers—that he acted honorably. We postulate that *A* cared about the public perception of his honorability, which was immediately manifest in the astonished reactions of the host and the studio audience. Viewers at that time, and later on YouTube, generally applauded his daring choice (there have been some 10 million views of the video and 10 thousand comments on it).

It was daring because it was risky and might well have backfired. So how did *A* prevent this when *B* has a strong incentive to Steal and possibly win the entire jackpot? He accomplished this by (i) promising that he would Steal so that *B* does not think that he can Steal himself and win everything, and (ii) credibly promising to Split after play of the original game, giving *B* the hope that he might obtain half the jackpot, even though *A*'s promise to Split later is cheap talk.

If *A* is able to dissuade *B* from choosing Steal at the outset, then he will be in a position to honor his promise to Split if he wins the jackpot. But rather than Splitting privately after play of the original game, *A* can choose Split at the outset and gain acclaim not only for his beneficence but also for his brilliance in telling a forgivable lie to undermine the dominance of his strategy of Steal in Game 2, which induces *B* to choose Steal himself.

We assume that *A*'s preferences remain the same as in Game 2, evidenced by *B*'s (actually, "Ibrahim's" in the episode) response https://www.wnycstudios.org/podcasts/radiolab/segments/golden-rule)
in a Radiolab interview to a question about whether he would have chosen to Split knowing that *A* ("Nick") was also going to Split: "No, never." Indeed, "it was Steal or nothing, because he would rather that both of them walk away without money than be duped into choosing Split, only to have all the money taken by the other contestant:" (https://blogs.pugetsound.edu/econ/2017/11/09/heart-of-gold-game-theory-in-game-show-golden-balls/)."
Clearly, Ibrahim's status as well as money counted for him.

How does *A*'s interest in keeping his promise and acting honorably—especially publicly—change his preferences in Game 2 (see Game 3)? First, the Split-Split outcome becomes (4,3) rather than (3,3), making it better for *A* than the other (3,3) outcome in Game 2, because *A* demonstrates his willingness to Split at the outset rather than privately after play of the original game.

Game 3

			B	
		Split		Steal
	Split	(4,3)*		(2,2)
A	Steal, then Split (if possible)	(3,3)		(2,2)
	Steal, then don't Split (if possible)	(1,1)		(2,1)

The second change in *A*'s preferences is more dramatic: (4,1) in Game 2 becomes (1,1) in Game 3, because *A* thoroughly dishonors himself by breaking his promise to Split

when *B* does. To be sure, *A* wins the entire jackpot, but he evinced no interest in doing so at the price of reneging on his promise, which he never had to do by choosing Split at the outset.

How do these two changes in *A*'s preferences affect the strategy choices of the players in Game 3? Now *A* has a weakly dominant strategy of Split, and *B*'s best response is also to Split, yielding the unique Nash-equilibrium outcome of (4,3), which we have starred in Game 3. This is exactly what occurred in the TV game.

In effect, we believe, preserving his honor was more important to *A* than maximizing his monetary payoff. Anomalously, however, *A* upheld his honor by Splitting in the original game, breaking his promise to Steal. This relieved him of the need to keep his promise to Split later if *B* also Split, which would not have been immediately visible to the audience whose approbation he sought.

A's announcement succeeded not only in persuading *B* to Split but also, by choosing Split himself in the original game, providing a public display that his promise was not just cheap talk. Indeed, *A*'s choice of Split in the beginning spoke louder than his words, making him appear even more honorable—publicly rather than just privately (if he Split later)—while achieving the same end.

Conclusions

Game 3 demonstrates *A*'s astuteness in using a lie to escape the Prisoners' Dilemma inherent in Golden Balls. It helps to explain why he announced he would Steal in the original game: It enabled him credibly to cow *B* into "submission," after adding the sop that he would Split the jackpot he won after the show (this promise was not so credible).

But by Splitting at the outset, *A* could honor his promise to Split the jackpot publicly, without actually having to do so privately, so we never learn whether or not he was lying that he would have Split the jackpot in the end. Clearly, *A*'s announcement was not only brilliantly devious, but it also worked.

Game 3 also explains how *A* overcame the credibility problem, given that his pledge to Split was cheap talk and therefore unconvincing to *B*. As we show, A's announcement, to which the (shell-shocked) moderator did not object--he could have declared, "No private deals of any kind!"--changed *B*'s perception of the game to Game 3, in which it was now rational for him to Split.[6]

Finally, it is worth noting that *A*'s announcement has a precedent in the famous Bible story in which two women claimed maternity of a baby. When King Solomon announced that he would split the disputed baby in two if both women refused to give up their claims, he elicited from them responses that revealed who the true mother was, so there was no need to split the baby[7]—just as *A*'s announcement in Golden Balls enabled both contestants to escape the dilemma in Prisoners' Dilemma and leave empty-handed if they both Steal.

---

[6] Talwalkar (2012) reaches a similar conclusion, based on a different model.

[7] See Brams (2018, ch. 10) for a more robust action that Solomon might have taken to settle the women's dispute which, unlike the threat he reneged on, could be repeated in future disputes.

References

Brams, Steven J. (2018). *Divine Games: Game Theory and the Undecidability of a Superior Being*. Cambridge, MA: MIT Press.

Burton-Chellew, Maxwell N., and Stuart A. West (2012). "Correlates of Cooperation in a One-Shot High-Stakes Televised Prisoners' Dilemma." *PLoS ONE* 7(4): e33344.

Cornell University Course Blog (2012). "Split or Steal: An Analysis Using Game Theory." <https://blogs.cornell.edu/info2040/2012/09/21/split-or-steal-an-analysis-using-game-theory/> (accessed Nov. 20, 2019).

Hart, Einav (2010). "Steal the Show Payoff Effect on Accuracy of Behavior-Prediction in Real High-Stake Dilemmas." Department of Cognitive Science, The Hebrew University.

Nikolaev, Boris (2014) "Using Experiments and Media to Introduce Game Theory into the Principles Classroom." *Journal of Private Enterprise* 29: 149-160.

Talwalkar, Presh (2012). "How to beat the Prisoner's Dilemma in the TV game show Golden Balls." <https://mindyourdecisions.com/blog/2012/04/24/how-to-beat-the-prisoners-dilemma-in-the-tv-game-show-golden-balls/> (accessed Nov. 20, 2019).

Turmunkh, Uyanga, Martijn J. van den Assem, and Dennie van Dolder (2019). "Malleable Lies: Communication and Cooperation in a High Stakes TV Game Show." *Management Science* 65(10): 4795–4812.

van den Assem, Martijn J., Dennie van Dolder, and Richard H. Thaler (2012). "Split or Steal? Cooperative Behavior When the Stakes Are Large." *Management Science* 58(1): 2