


Malleable Lies: Communication and Cooperation in a High Stakes TV Game Show

Uyanga Turmunkh,^a Martijn J. van den Assem,^b Dennie van Dolder^{b,c}

^a Department of Economics and Quantitative Methods, IÉSEG School of Management, 59000 Lille, France; ^b School of Business and Economics, Vrije Universiteit Amsterdam, 1081 HV Amsterdam, Netherlands; ^c Centre for Decision Research and Experimental Economics, University of Nottingham, Nottingham NG7 2RD, United Kingdom

Contact: u.turmunkh@ieseg.fr,  <http://orcid.org/0000-0003-3623-0564> (UT); m.j.vanden.assem@vu.nl,  <http://orcid.org/0000-0001-5166-1288> (MJvdA); d.van.dolder@vu.nl,  <http://orcid.org/0000-0002-2498-8311> (DvD)

Received: January 29, 2017

Revised: May 3, 2018

Accepted: June 15, 2018

Published Online in Articles in Advance:
March 27, 2019

<https://doi.org/10.1287/mnsc.2018.3159>

Copyright: © 2019 INFORMS

Abstract. We investigate the credibility of nonbinding preplay statements about cooperative behavior, using data from a high-stakes TV game show in which contestants play a variant on the classic Prisoner's Dilemma. We depart from the conventional binary approach of classifying statements as promises or not, and propose a more fine-grained two-by-two typology inspired by the idea that lying aversion leads defectors to prefer statements that are malleable to ex-post interpretation as truths. Our empirical analysis shows that statements that carry an element of conditionality or implicitness are associated with a lower likelihood of cooperation, and confirms that malleability is a good criterion for judging the credibility of cheap talk.

History: Accepted by Elke Weber, judgment and decision making.

Funding: This work was supported by the Economic and Social Research Council via the Network for Integrated Behavioural Sciences [Grant ES/K002201/1] and the Nederlandse Organisatie voor Wetenschappelijk Onderzoek [Grant 452-16-011].

Supplemental Material: Data are available at <https://doi.org/10.1287/mnsc.2018.3159>.

Keywords: deception • lying aversion • game show • prisoner's dilemma • communication • cooperation • cheap talk

1. Introduction

In many social and economic interactions, people care about the behavior of others. For instance, in fiduciary business relationships, principals are concerned about the risk of being exploited by their agents, and in romantic relationships, lovers considering moving in together may worry about the eventual division of household labor. Often the information that is available for predicting what others will do is limited to non-binding and nonverifiable communication.

The traditional assumption in economics is that lying is costless. Under that assumption, cheap talk is not informative when interests are insufficiently aligned (Farrell and Rabin 1996). This assumption, however, contradicts the view held in social psychology that for most people lying entails an unpleasant and therefore costly experience (Ekman 2001, Vrij 2008). An extensive body of research on cues to deception critically builds on the premise that lying and truth-telling generate distinct emotional states (Zuckerman et al. 1981, DePaulo et al. 2003, Sporer and Schwandt 2007). Economic experiments indeed reject the notion that people are unscrupulous liars and show that cheap talk can be informative in the absence of aligned interests (Gneezy 2005, Lundquist et al. 2009, Serra-Garcia et al. 2011, Erat and Gneezy 2012, Cappelen et al. 2013, Abeler et al. 2014).

There are a number of explanations for the aversion to lying. Commitment-based explanations posit that people simply have an intrinsic preference for keeping their word (Ellingsen and Johannesson 2004, Vanberg 2008). Such a preference is also in line with self-concept maintenance theory, where people like to view themselves as honest and are averse to negatively updating their self-concept after a dishonest act (Mazar et al. 2008). Alternatively, expectation-based accounts state that people dislike lying because they experience guilt if they do not live up to others' expectations (Charness and Dufwenberg 2006, 2010; Battigalli et al. 2013; Ederer and Stremitzer 2017). People may also refrain from lying simply because of fear of being caught, which could harm their reputation as an honest and reliable person.¹

The literature on lying aversion and the predictive power of promises has largely concentrated on binary—to lie or not to lie—stylized choice contexts. The present paper extends this literature with a more fine-grained conceptual framework. When talk is free-form, as in most real-life situations, the set of possible deceptive statements is richer: instead of outright lying, people can choose to deceive by omitting, obfuscating, or stretching the truth. This wider choice set is potentially important, because some types of deception may be more aversive to liars than other types and thus

more likely to be avoided. As an implication of this avoidance, some statements will be more indicative of cooperative behavior than others. Although some papers do study the predictive power of statements in free-form communication, they typically distinguish only one particular type of message—a promise—so that cheap talk is effectively still analyzed in a binary framework (Vanberg 2008, van den Assem et al. 2012).

In the present paper we empirically address the question whether distinguishing between different types of statements adds to the predictive power of cheap talk, using a high-stakes noncooperative game with free-form communication. In each episode of the British TV show *Golden Balls*, two contestants play a variant on the Prisoner's Dilemma where they simultaneously decide to either split (cooperate) or steal (defect) a sum of money that on average exceeds £13,000. Prior to their decisions, they engage in a brief free-form discussion about the choice at hand. During the talk, they typically exchange multiple statements, most of which involve giving or eliciting some type of signal that the intended decision is to split.²

We hypothesize that lies are less costly if they are malleable to interpretation as truths and that people who defect prefer statements that allow them to deny—to themselves or to others—that they are lying. Such statements arguably entail weaker commitment, weaker feelings of disappointing the other, and lower reputation costs. People who cooperate have no reason to resort to malleable statements. If defectors avoid unmalleable statements and cooperators do not, malleability becomes a criterion for judging the credibility of cheap talk.

Earlier studies providing support for the idea that people want to avoid blatant lies when they try to mislead others include Serra-Garcia et al. (2011) and Khalmetski et al. (2017). Serra-Garcia et al. find that people rather deceive by means of a vague message that captures the truth. Khalmetski et al. demonstrate that people often engage in evasive lying by pretending not to know the truth. Furthermore, a growing line of work shows that people are more willing to cheat if the context provides some leeway to justify their behavior (Schweitzer and Hsee 2002, Mazar et al. 2008, Shalvi et al. 2011, Pittarello et al. 2015).³

An implication of our malleability hypothesis is that unmalleable statements or “promises” become less predictive of cooperation when defectors cannot resort to more malleable statements. Empirical work by Charness and Dufwenberg (2006, 2010) on free-form versus restricted communication and by Belot et al. (2010) on voluntary versus elicited promises supports this prediction.

We propose a typology of statements in terms of their malleability to interpretation as truths. This typology classifies contestants' statements according to two dimensions. First, it discriminates between statements

that explicitly express that the contestant will choose split and statements that only implicitly signal that she will do so. Second, it discriminates between unconditional statements and statements that carry an element of conditionality on the opponent's split or steal decision. We argue that explicit and unconditional statements are less malleable than implicit or conditional statements. Consider, for example, the statement “I will split.” This statement is both explicit and unconditional, and for a defector who uses it, it will be hard if not impossible to deny that she has deceived her opponent. The statement “I came here to split” similarly has no element of conditionality, but this one is at best only an implicit promise to split: it is silent about the contestant's current intention and meanwhile she may have changed her mind. The explicit statement “I will split if you split” is clearly conditional on the opponent's choice, and a decision to steal can be justified by a belief that the opponent steals.

Our empirical analysis confirms that malleability is a good criterion for judging the credibility of cheap talk. Explicit unconditional statements are indicative of a relatively high likelihood of cooperation, whereas statements that carry an element of conditionality or implicitness are associated with a moderate likelihood. Contestants who make statements that are both conditional and implicit and contestants who do not make any statements related to their choice display the lowest rate of cooperation. In spite of this predictive power and in spite of the evidence in the literature that people have conditionally cooperative preferences (Fischbacher et al. 2001, Frey and Meier 2004), we find no evidence that contestants condition their choice on their opponents' statements.

2. Game Show and Data

2.1. *Golden Balls*

Golden Balls was broadcast on British television between June 2007 and December 2009. We analyze the game played in the fourth and final round of every episode. In this final, two contestants play a game that resembles the classic Prisoner's Dilemma. What follows here is a brief description of *Golden Balls*. For a more extensive discussion we refer to van den Assem et al. (2012).

The show begins with four contestants who have not met before. In the first two rounds, the four (round 1) or remaining three (round 2) contestants all receive a set of golden balls. Each ball carries a specific value, which in the end may contribute to the final jackpot. The contents of some balls are visible to everyone, whereas the contents of the other balls are known to the contestant only. Contestants have to make claims about the contents of their hidden balls, after which they have an open discussion and then cast votes against each other. The player who receives most votes is eliminated from

the game, together with her golden balls. In the third round, the two remaining contestants determine the jackpot through a random draw from the remaining balls.

In the fourth and final round, each of the two finalists is presented with two golden balls, one with the word “split” and the other with the word “steal” written inside. They simultaneously have to choose either the split or the steal ball. If both decide to split, they split the jackpot equally. If one decides to split while the other decides to steal, the one who steals receives the entire jackpot and the one who splits goes home with nothing. If both decide to steal, both go home with nothing. Following Rapoport’s (1988) terminology, the choice problem can be labeled as a “weak” form of the Prisoner’s Dilemma because defection does not strictly dominate cooperation. Prior to their decisions, the two contestants engage in a brief free-form discussion in which they can try to persuade one another to cooperate.

2.2. Van den Assem et al. (2012)

Golden Balls has previously been studied by van den Assem et al. (2012). The analyses in the present paper control for the determinants of cooperation identified in that study. Here we summarize the earlier findings.

Young males are less cooperative than young females. This gender difference reverses for older contestants because older men are much more cooperative than younger men. In addition, there is some evidence that white and higher educated contestants are more likely to cooperate. Distinguishing between contestants from urban and rural areas and between students and nonstudents adds little explanatory power.

Cooperation is largely insensitive to the size of the jackpot: contestants cooperate about half the time, irrespective of whether they are playing for a few thousand or a hundred thousand pounds. A major exception is that the rate of cooperation is relatively high, about 70%, when there are only a few hundred pounds at stake.

During the show’s first seasons—when few or no episodes had been broadcast and contestants were not yet able to accurately estimate what they could expect to win in the show—the choices of contestants were strongly influenced by the maximum potential jackpot prior to the final. Before players determine the jackpot by randomly drawing five golden balls, a great deal of attention is given to this maximum. The higher this maximum is—and the smaller the actual jackpot thus appears to be—the greater the likelihood that players cooperate. This effect diminishes with the number of televised episodes, possibly because players learned what to expect. These results suggest that contestants assess the stakes in relative terms.

Contestants who misrepresented the contents of their hidden balls in the first two rounds of the game do not cooperate more or less than those who have been honest throughout, and neither do their opponents. Consistent with the notion that people have a preference for reciprocity, contestants are less likely to cooperate with opponents who have tried to vote them off the show during the first two rounds.

The prior study also analyzes the predictive power of promises in a binary framework and reports that those who make an unambiguous promise are much more likely to cooperate. Unambiguous promises resemble what we call explicit unconditional statements in the present paper, with the important difference that we now ignore statements that have no meaning on their own, such as short utterances in response to remarks or questions by the opponent (e.g., “Yeah,” “No,” and “Absolutely”; see Section 2.4).⁴

Last, van den Assem et al. (2012) find little evidence that contestants’ propensity to cooperate depends on the likelihood that their opponent cooperates: players do not seem to condition on their opponent’s promise or background characteristics, despite their predictive power.

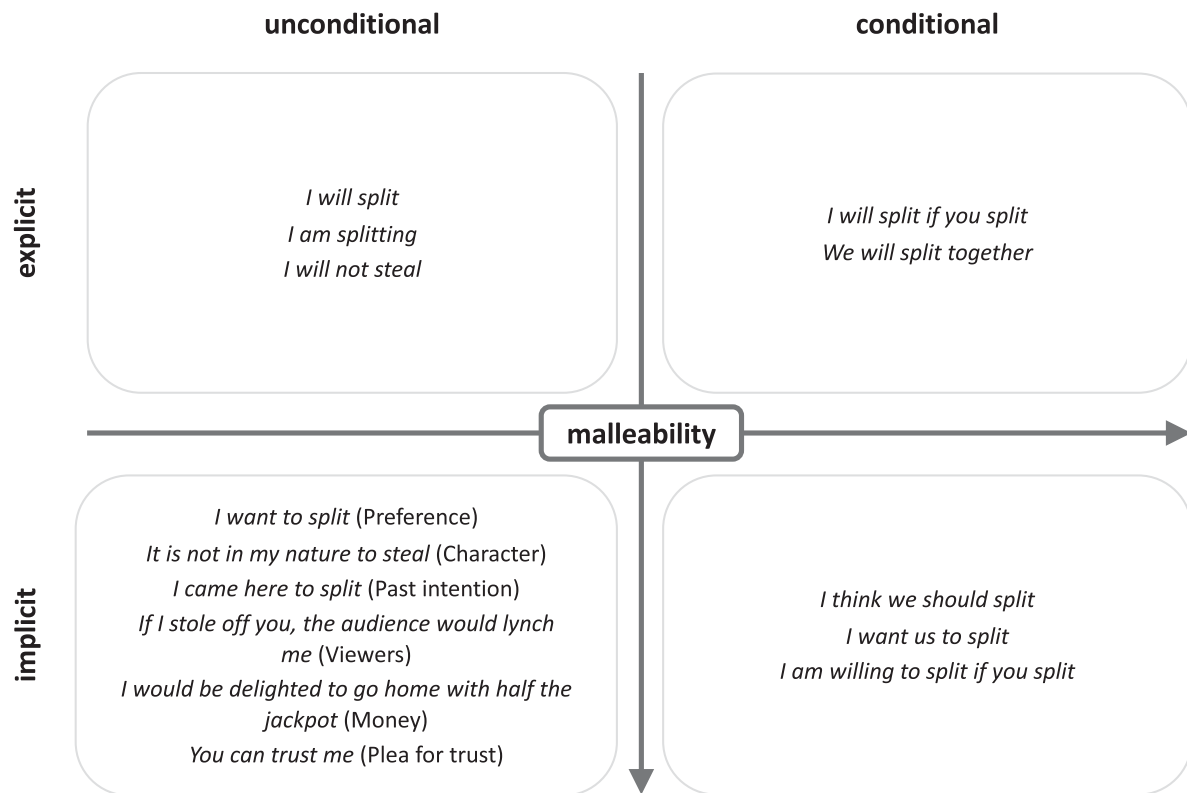
2.3. Typology of Preplay Communication

Prior to their split or steal decisions, the two finalists engage in a brief free-form talk about the choice at hand. During this talk, they typically exchange multiple statements through which they try to signal that they will choose split or to get assurances that the opponent will do so. Owing to the free-form nature, there is a rich set of statements that contestants can make. In this subsection, we introduce the typology of statements that is central in the present paper.

We categorize the different statements that signal splitting according to how malleable they are. Contestants who are about to steal normally try to conceal this intention. On the basis of the notion that lies that are more malleable to interpretation as truths entail lower lying costs, we expect that stealing contestants will prefer malleable statements over unmalleable statements. Contestants who are about to split face no lying costs and thus have no need for such substitution. As a consequence, the malleability of the statements that contestants make will predict cooperative behavior.

Our typology distinguishes between explicit and implicit statements, and between unconditional and conditional statements. These distinctions lead to four different categories (Figure 1).

Explicit unconditional statements such as “I will split” are completely unmalleable. Consider, in contrast, the statement “I want to split.” The latter does not explicitly express that the contestant will choose split, but signals this by expressing a preference to do so.

Figure 1. Typology of Statements

Notes. The typology classifies all statements that signal that the contestant will cooperate into one of four categories on the basis of whether they explicitly or only implicitly say that the individual will cooperate, and on the basis of whether they carry an element of conditionality on the opponent's split or steal decision. Implicit unconditional statements are divided into six different subcategories.

Strictly speaking, the desire to split can be sincere even if the action is to the contrary, and therefore we argue that this statement is more malleable than "I will split." In a similar way, a contestant can signal that she will split without stating so explicitly by referring to her character ("It is not in my nature to steal"), by referring to a past intention ("I came here to split"), by expressing awareness or worry that viewers will disapprove stealing ("If I stole off you, the audience would lynch me"), by indicating that half the jackpot is good enough or a lot of money ("I would be delighted to go home with half the jackpot"), or by urging the opponent to have trust ("You can trust me"). All these types of implicit unconditional statements allow a contestant to deny that she has lied if she steals.

Next, compare the explicit unconditional statement "I will split" with the statements "I will split if you split" and "We will split together." All are explicit about what the contestant will do, but the latter two carry an element of conditionality on the opponent's split or steal decision: through an if-clause or because it refers to the coordinated action. In the latter case the element of conditionality is not as literally present as it is in the former, but in both cases the contestant is connecting her own choice with that of her opponent. The conditionality in explicit conditional statements

allows a stealing contestant to deny that she has lied, as she can always argue that she did not believe that her opponent was about to split.⁵

The fourth category comprises statements that are both implicit and conditional. These two properties make it easy to deny lying if one decides to steal. Examples include "I think we should split," "I want us to split," and "I am willing to choose split if you split."

Our typology only considers statements that signal that the contestant will choose split. All other things contestants say are considered as empty talk. Empty talk includes questions and pleas directed at the opponent ("Will you split?" or "I hope you will split"), remarks about the previous rounds ("I did not lie a single time" or "I brought you to the final"), and many other types of idle remarks ("This is a lot of money," "You seem like a nice person," or "We have come so far"). Additional examples of statements and explanations of how they are coded are in the appendix.

2.4. Coding Rules

Coding free-form communication into the types described above is not straightforward. People stumble over words, jump from one topic to the next, and use short and incomplete phrases that are difficult to interpret. We use the following strict set of coding rules.

First, we require that statements have a meaning on their own. This implies that we ignore short utterances such as “Yeah,” “No,” and “Absolutely.” Short utterances are often preceded or followed by complete statements that elaborate their meaning. Complete statements are coded, so in those cases no information is lost. Furthermore, many short utterances constitute responses to remarks or questions by the opponent. In a conversation between two people, people are more or less expected to respond in a particular way to what the other has said or asked. If a contestant chooses not to back up her short response with a longer statement, this can be seen as an indication that the short reply was forced out of her. In this sense, our analysis thus only considers statements that contestants have made voluntarily.⁶

Second, we count consecutive statements of the same type as one statement if they are not interspersed with other statements. Contestants frequently repeat what they say or elaborate on it (“It is not in my nature to steal. I just could not do that.”). Such repetitions and elaborations are hard to count and separate, and combined they represent a single message. When multiple statements of the same type are interspersed with other statements, we do treat them separately, as this indicates that the contestant consciously decided to make the same point again.

Third, if the stand-alone interpretation of the verb “to split” is ambiguous, we assume that it refers to the contestant’s action of choosing the split ball. Formally, the verb can refer both to the action of a single contestant who chooses the split ball and to the outcome of two contestants actually sharing the jackpot. Often the exact meaning is clear, but for some statements, such as “I want to split,” one can argue both ways. Because the more common meaning in the show is the unilateral act of picking the split ball (the game of interest is even called “Split or Steal?”), we use this meaning as the default interpretation. For consistency, we treat the synonym “to share” analogously.

Fourth, just like “you and I” is equal to “we,” we interpret the combination of a statement referring to “you” (the opponent) and a similar statement referring to “I” (the player herself) as one statement referring to “we.” For example, “You will split, I will split” is considered equivalent to “we will split.” The order of the two statements does not matter, but we do require that they directly follow each other.

Last, if a statement contains a pronoun such as “it,” “this,” or “that” and if the noun to which it refers is clear from the context, then we interpret such a statement as if that noun replaces the pronoun.

2.5. Data and Descriptive Statistics

The data set used for this study covers 284 episodes. Recordings were originally provided by Endemol’s

local production company Endemol UK for the purpose of the study by van den Assem et al. (2012). Of the 288 episodes that aired in total, one could not be located by the producer and is therefore missing. We excluded three more episodes. In one of these, a finalist announces that he will pick the steal ball, and promises to reward his opponent with half of the money after the show if the opponent splits. This exceptional case is incompatible with our coding framework.⁷ In the two other excluded episodes, most of what contestants say is clearly meant ironically, probably owing to the small jackpot sizes (£3.00 and £3.65).

The 568 contestants in our sample are playing for a jackpot that ranges between a few pounds and about one hundred thousand pounds, and averages £13,510 (median: £4,325). Approximately half (52%) cooperate. The average age is 37 years (median: 34 years), and there are about as many males (46%) as females (54%).

To construct our communication data, we first transcribed all conversations. For each contestant, the three authors then independently counted the number of statements in every category. Differences were discussed until consensus was reached.

Table 1 shows the frequency distribution of the number of statements made by contestants for each statement category. Over one-third of the contestants (38%) made at least one explicit unconditional statement (“I will split”). More than half (56%) made one or more implicit unconditional statements. The most frequent type of implicit unconditional statement relates to the jackpot size (25%) or refers to a preference (19%) or past intention (13%). Relatively few contestants (7%) made an explicit conditional statement (“I will split if you split” or “We will split together”), and about one-third (34%) made an implicit conditional one (“I am willing to choose split if you split” or “I think we should split”).⁸ The use of multiple statements from the same category is relatively rare. About one in six (17%) limited themselves to empty talk.

3. Analyses and Results

3.1. Preliminary Analysis

For a first glimpse of the association between the malleability of statements and cooperation, we consider the rates of cooperation after different statement types. Figure 2(a) shows the cooperation rate conditional on whether a contestant made a statement from a particular category, regardless of repetitions and regardless of whether she also made statements from any other category. In line with our malleability hypothesis, the cooperation rate is highest for contestants who made an explicit unconditional statement (67%). The rate is lower after the more malleable implicit unconditional (58%) and explicit conditional (59%) statements. Contestants who made a statement that is

Table 1. Frequency Distributions of the Number of Statements

Statement category	Frequency of statements					
	0	>0	1	2	3	>3
Explicit unconditional	353 (62.1%)	215 (37.9%)	146 (25.7%)	54 (9.5%)	11 (1.9%)	4 (0.7%)
Implicit unconditional	250 (44.0%)	318 (56.0%)	184 (32.4%)	95 (16.7%)	26 (4.6%)	13 (2.3%)
Character	532 (93.7%)	36 (6.3%)	31 (5.5%)	5 (0.9%)	0 (0.0%)	0 (0.0%)
Past intention	493 (86.8%)	75 (13.2%)	63 (11.1%)	11 (1.9%)	0 (0.0%)	1 (0.2%)
Viewers	534 (94%)	34 (6.0%)	31 (5.5%)	3 (0.5%)	0 (0.0%)	0 (0.0%)
Preference	462 (81.3%)	106 (18.7%)	90 (15.8%)	12 (2.1%)	4 (0.7%)	0 (0.0%)
Trust	516 (90.8%)	52 (9.2%)	46 (8.1%)	6 (1.1%)	0 (0.0%)	0 (0.0%)
Money	427 (75.2%)	141 (24.8%)	123 (21.7%)	18 (3.2%)	0 (0.0%)	0 (0.0%)
Explicit conditional	529 (93.1%)	39 (6.9%)	35 (6.2%)	3 (0.5%)	1 (0.2%)	0 (0.0%)
Implicit conditional	376 (66.2%)	192 (33.8%)	153 (26.9%)	32 (5.6%)	7 (1.2%)	0 (0.0%)
Overall	96 (16.9%)	472 (83.1%)	162 (28.5%)	150 (26.4%)	76 (13.4%)	84 (14.8%)

Notes. This table shows the frequency distribution of the number of statements made by contestants for each statement category and for the subcategories of implicit unconditional statements. Percentages of the total pool of 568 contestants are in parentheses.

both implicit and conditional cooperate even less (49%). Last, contestants who used none of the four types have the lowest propensity to split (31%).

Contestants normally make more than one statement. Figure 2(b) shows that there are no clear incremental effects of one or more additional statements from the same category, suggesting that the propensity to split is similar for contestants who made a particular statement only once and for those who made the same statement multiple times.

Many make statements from different categories, and the malleability of a set of statements may be fully driven by the strongest statement alone. For example, a contestant who explicitly promises her opponent that she will split arguably does not make her overall message more or less malleable by adding that viewers would lynch her if she steals.

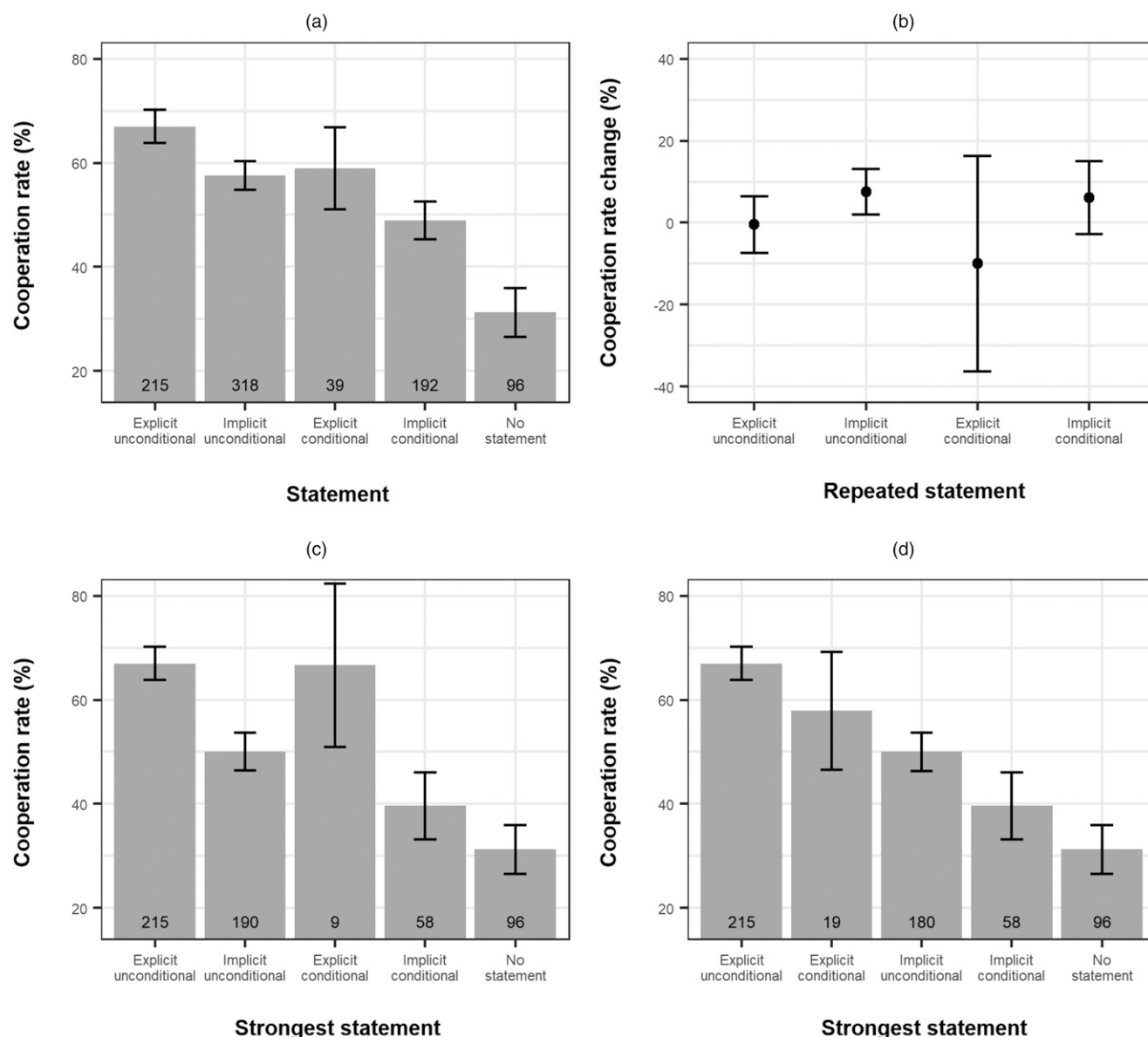
Figures 2(c) and 2(d) indicate how cooperation depends on a contestant's strongest statement. We use two different orderings, because there is no compelling a priori reason to consider implicit unconditional statements more or less malleable than explicit conditional statements. Figure 2(c) assumes that implicit unconditional statements are stronger than explicit conditional statements, and Figure 2(d) assumes the reverse. The patterns are consistent with the malleability hypothesis. The cooperation rate is highest when

the strongest statement was of the explicit unconditional type (67%), low when the strongest was both implicit and conditional (40%), and lowest when the contestant limited herself to empty talk (31%). Under both orderings, contestants whose strongest statement was implicit unconditional or explicit conditional show intermediate cooperation rates. The averages suggest that explicit conditional statements are stronger than implicit unconditional ones [as assumed in Figure 2(d)].

3.2. Regression Analysis

The comparisons of cooperation rates provide only crude insights into the predictive power of communication, because they neither take into account the differences in contestant and game characteristics nor the use of multiple statements by the same contestant. In this section, we therefore turn to multivariate Logit regressions. The dependent variable is the contestant's decision, taking the value of one for "split" and zero for "steal." The main regressors are variables for the different statement categories. The set of control variables is identical to the set of variables in model 6 of van den Assem et al. (2012, p. 13). It includes demographic information on age, gender, race, place of residence, and education, and game characteristics describing the jackpot, how many times the show

Figure 2. Cooperation Across Statement Categories



Notes. Panel (a) displays the percentage of cooperators for all contestants who made at least one statement from a particular category and for all contestants who limited themselves to empty talk. Panel (b) displays the differences in the percentages of cooperators between contestants who made at least two statements from a particular category and those who made only one. Panels (c) and (d) display the percentage of cooperators for all contestants whose strongest statement belongs to a particular category, with panel (c) assuming that implicit unconditional statements are stronger than explicit conditional statements, and panel (d) assuming the reverse. The number of contestants is at the bottom of each bar. Error bars depict standard errors around the mean.

had been aired prior to recording and whether the opponent has tried to vote the participant off the show. We follow the common approach of reporting average marginal effects (with the corresponding standard errors and significance levels) and correct the standard errors for clustering at the episode level (Wooldridge 2003).

In Table 2, Models 1 and 2 use indicator variables for the statements and ignore possible repetitions. Model 2 includes the full set of controls and shows that a contestant who made one or more explicit

unconditional statements is 25 percentage points more likely to choose split than a contestant who did not make any such statements ($p < 0.001$). Implicit unconditional statements also predict a higher likelihood of splitting, but to a lesser extent: the average marginal effect is 12 percentage points ($p = 0.002$). Explicit conditional and implicit conditional statements are insignificant predictors of behavior.

The marginal effects of the four statement types differ significantly (Wald $\chi^2(3) = 26.58$, $p < 0.001$), and rank in line with the malleability hypothesis.

Table 2. Regression Results for Statement Indicators and Proportions

	Model 1	Model 2	Model 3		Model 4	Model 5
	<i>One or more</i>	<i>One or more</i>	<i>One or more</i>	<i>Two or more</i>	<i>Proportion</i>	<i>Proportion</i>
Explicit unconditional	0.224*** (0.043)	0.253*** (0.041)	0.242*** (0.046)	0.021 (0.061)	0.403*** (0.057)	0.426*** (0.051)
Implicit unconditional	0.101** (0.042)	0.122*** (0.040)	0.091** (0.044)	0.074 (0.052)	0.214*** (0.046)	0.233*** (0.041)
Explicit conditional	0.038 (0.082)	0.022 (0.082)	0.016 (0.087)	0.025 (0.168)	0.259* (0.136)	0.277** (0.109)
Implicit conditional	−0.036 (0.044)	−0.019 (0.043)	−0.027 (0.046)	0.062 (0.088)	0.073 (0.063)	0.103* (0.057)
Controls	No	Yes	Yes		No	Yes
McFadden R^2	0.047	0.136	0.139		0.054	0.141
Number of clusters	284	284	284		284	284
Observations	568	568	568		568	568

Notes. The table reports the average marginal effects resulting from Logit regression analyses of contestants' decisions to split (1) or steal (0) the jackpot. Models 1 and 2 use indicator variables measuring whether the contestant made at least one statement that belongs to the given category. Model 3 adds variables indicating whether the contestant made two or more statements belonging to the given category. Models 4 and 5 employ variables measuring the proportion of the contestant's statements that belong to the given category. The set of control variables in Models 2, 3, and 5 is identical to the variables in model 6 of van den Assem et al. (2012, p. 13). It includes demographic information on age, gender, race, place of residence, and education, and game characteristics describing the jackpot, how many times the show had been aired prior to recording, and whether the opponent has tried to vote the participant off the show. Standard errors (in parentheses) are corrected for clustering at the episode level.

* $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Pairwise comparisons provide further evidence for the malleability hypothesis: the marginal effect of explicit unconditional statements is significantly larger than that of each of the three more malleable types of statements (all Wald $\chi^2(1) > 4.89$, all one-sided $p < 0.013$), and the marginal effect of implicit unconditional statements is larger than that of implicit conditional statements (Wald $\chi^2(1) = 5.80$; one-sided $p = 0.008$). There is no significant difference between explicit conditional and implicit conditional statements (Wald $\chi^2(1) = 0.18$; one-sided $p = 0.335$). The malleability hypothesis makes no prediction regarding the relative strength of implicit unconditional and explicit conditional statements; statistically the marginal effects of these statement types do not differ (Wald $\chi^2(1) = 1.21$; $p = 0.272$).

Existing research into the credibility of cheap talk typically uses a binary framework where people either make a promise or not. When promises are equated to explicit unconditional statements, implicit statements and conditional statements do not exist by design or are categorized as empty talk. Our more fine-grained two-by-two typology improves the explanatory power over a model that only considers explicit unconditional statements and pools the other three types with empty talk: the null hypothesis that the coefficients for implicit unconditional, explicit conditional and implicit conditional statements are jointly equal to zero can be rejected (Wald $\chi^2(3) = 9.19$, $p = 0.027$). Adding the three weaker statement types to the model with explicit unconditional statements and the set of controls increases the McFadden R^2 from 0.123 to 0.136 and the

"hit" percentage—the fraction of correctly predicted decisions—from 64.8 to 67.3.⁹

Model 3 tests whether there are incremental effects of repeating statements of the same type. As indicated by the earlier preliminary analysis, the predictive power fully derives from whether a particular statement has been made, and not from possible repetitions. All repetition variables are insignificant and combined the four add no significant explanatory power to the model (Wald $\chi^2(4) = 2.76$, $p = 0.598$). The marginal effects of the indicator variables for whether a statement occurred at least once are barely affected by the inclusion of indicators for repetitions.

In Models 4 and 5, the variables of interest measure the proportion of a contestant's statements that belong to the given category. This alternative approach assumes that the importance of a particular statement for the contestant's overall message depends on the number of other statements made. Model 5 includes the full set of controls and shows that a contestant who made explicit unconditional statements only is 43 percentage points more likely to cooperate than a contestant who used none of the four statement types ($p < 0.001$). Contestants who made implicit unconditional statements only or explicit conditional statements only are, respectively, 23 percentage points and 28 percentage points more likely to cooperate (both $p < 0.011$). Those who made implicit conditional statements only are 10 percentage points more likely to cooperate, but this effect is only marginally significant ($p = 0.072$).

The marginal effects differ significantly between the four statement types (Wald $\chi^2(3) = 18.01$, $p < 0.001$), and their ranking supports the malleability hypothesis. Pairwise tests show that explicit unconditional statements have a larger marginal effect than both implicit unconditional and implicit conditional statements (both Wald $\chi^2(1) > 8.18$, both one-sided $p < 0.003$) and that implicit unconditional statements have a stronger marginal effect than implicit conditional ones (Wald $\chi^2(1) = 4.55$, one-sided $p = 0.016$). There is no significant difference between explicit unconditional and explicit conditional statements (Wald $\chi^2(1) = 1.55$, one-sided $p = 0.106$) and a marginally significant difference between explicit and implicit conditional statements (Wald $\chi^2(1) = 2.24$, one-sided $p = 0.067$).

The null hypothesis that the coefficients for implicit unconditional, explicit conditional and implicit conditional statements are jointly equal to zero is rejected (Wald $\chi^2(3) = 18.41$, $p < 0.001$). For a clean measurement of the improvement of the empirical fit from distinguishing malleable statements from empty talk, however, Model 5 formally cannot be compared with a model that omits the three weaker types because the proportion of explicit unconditional statements by construction is the complement of the proportion of weaker statements. Using the model with an indicator variable for explicit unconditional statements as the benchmark model, Model 5 increases the McFadden R^2 from 0.123 to 0.141 and the hit percentage from 64.8 to 67.3.

Table 3 shows the marginal effects for contestants' strongest statements. Panel A assumes that implicit unconditional statements are stronger than explicit conditional statements, whereas panel B assumes the reverse. Accounting for controls, explicit unconditional statements increase the likelihood of cooperation by approximately 40 percentage points ($p < 0.001$). Explicit conditional and implicit unconditional statements have intermediate effects between 21 and 41 percentage points (all $p < 0.011$). Statements that are malleable because they are both implicit and conditional have no significant effect ($p > 0.177$).

For both rankings, the marginal effects of the four strongest-statement categories are significantly different (both Wald $\chi^2(3) > 26.38$, both $p < 0.001$), and their orders in terms of size again confirm the hypothesis that more malleable statements are less predictive of cooperative behavior. Pairwise tests show that the marginal effect of explicit unconditional statements is significantly larger than that of implicit unconditional and implicit conditional statements (all Wald $\chi^2(1) > 16.10$, all one-sided $p < 0.001$). Implicit unconditional statements are marginally stronger than implicit conditional statements (Model 2: Wald $\chi^2(1) = 2.63$, one-sided $p = 0.053$; Model 4: Wald $\chi^2(1) = 2.79$, one-sided $p = 0.047$). Explicit conditional statements are not different from explicit unconditional statements (both

Table 3. Regression Results for the Strongest Statement

Panel A: Implicit unconditional > explicit conditional		
	Model 1	Model 2
1. Explicit unconditional	0.357*** (0.059)	0.400*** (0.055)
2. Implicit unconditional	0.188*** (0.059)	0.214*** (0.056)
3. Explicit conditional	0.354** (0.165)	0.406*** (0.122)
4. Implicit conditional	0.084 (0.081)	0.103 (0.076)
Controls	No	Yes
McFadden R^2	0.053	0.145
Number of clusters	284	284
Observations	568	568
Panel B: Explicit conditional > implicit unconditional		
	Model 3	Model 4
1. Explicit unconditional	0.357*** (0.059)	0.400*** (0.055)
2. Explicit conditional	0.266** (0.117)	0.285** (0.111)
3. Implicit unconditional	0.188*** (0.060)	0.217*** (0.056)
4. Implicit conditional	0.084 (0.081)	0.102 (0.076)
Controls	No	Yes
McFadden R^2	0.052	0.143
Number of clusters	284	284
Observations	568	568

Notes. The table reports the average marginal effects resulting from Logit regression analyses of contestants' decisions to split (1) or steal (0) the jackpot. The models use indicator variables for the strongest statement that the contestant made. Models 1 and 2 in panel A assume that implicit unconditional statements are stronger than explicit conditional statements, and Models 3 and 4 in panel B assume the reverse. Other definitions are as in Table 2.

* $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Wald $\chi^2(1) < 1.15$, both one-sided $p > 0.141$), but more predictive of cooperation than implicit conditional statements (Model 2: Wald $\chi^2(1) = 6.26$, one-sided $p = 0.006$; Model 4: Wald $\chi^2(1) = 2.53$, one-sided $p = 0.056$). However, the latter test results need to be interpreted with some caution because there are only a handful of contestants whose strongest statement is of the explicit conditional statement type (9 or 19, depending on the ranking).

Just as in the analyses with indicator and proportion variables, our more fine-grained two-by-two typology also improves the explanatory power over a simple binary approach when contestants' strongest statements are considered. The null hypothesis that the coefficients of the more malleable strongest-statement variables are jointly equal to zero is rejected (both Wald $\chi^2(3) > 14.64$, both $p < 0.003$). Incorporating them increases the McFadden R^2 from 0.123 to 0.145 (Model 2)

or 0.143 (Model 4) and increases the hit percentage by from 64.8 to 66.7 (for both models).

Overall, the regression results support the typology introduced in Section 2.3, regardless of how we construct the statement variables. In addition to these main analyses we have examined the following potential moderator variables for the relation between statements and behavior (not tabulated). First, we considered the moderating effect of gender. Recent findings suggest that men are more willing to lie than women.¹⁰ In our data, there is no evidence of such a gender difference: the average marginal effects of statements do not differ significantly between males and females. Second, we investigated whether the relationships between statements and cooperation change with the number of episodes aired prior to the day of recording. Similar to the stable degree of cooperation over time found earlier (van den Assem et al. 2012), contestants appear to remain equally likely to lie: all interactions of statement variables and the number of past transmissions are statistically insignificant. Last, we looked at the moderating effect of jackpot size.

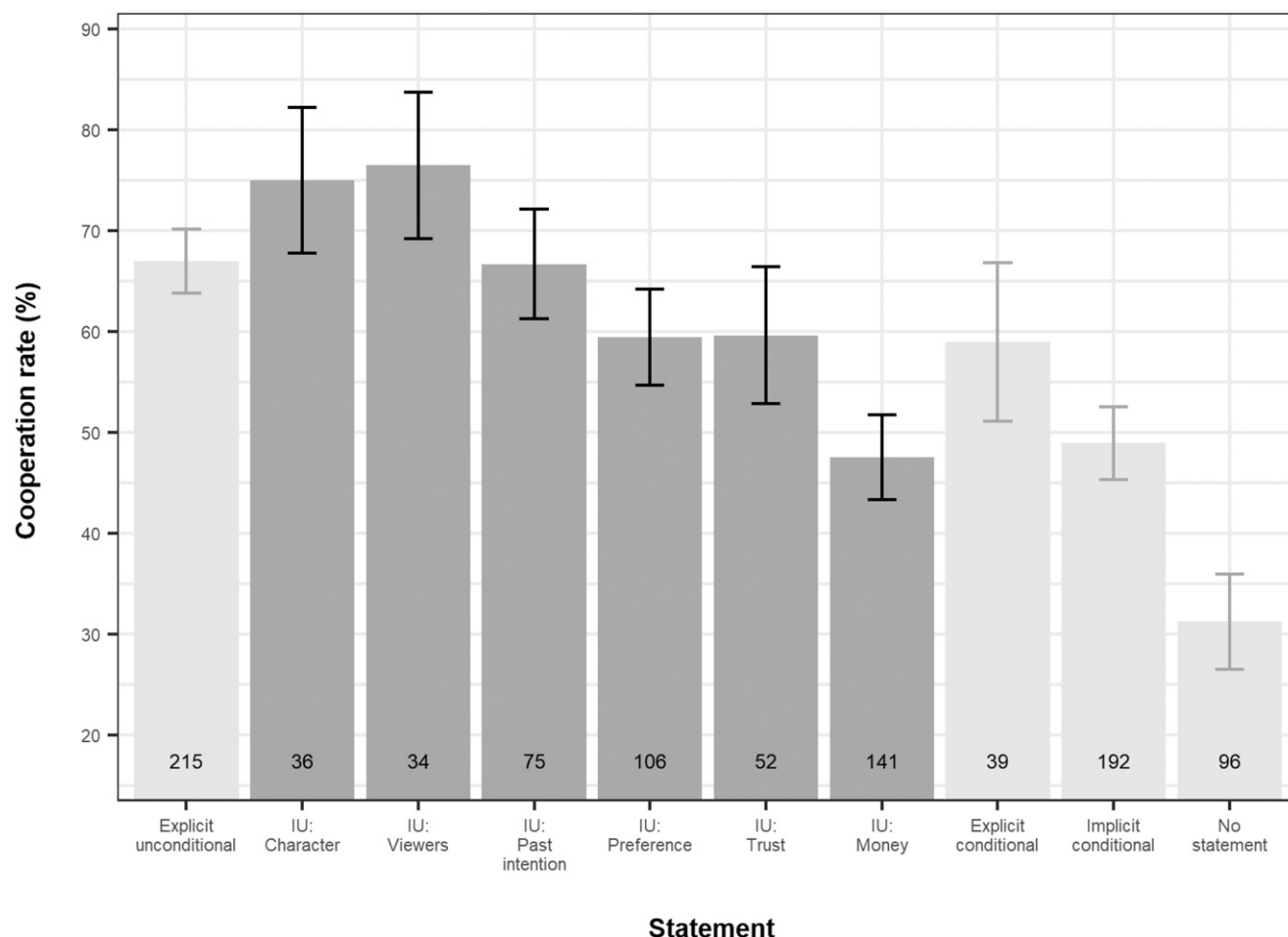
A meta-analysis of experiments employing stakes up to tens of dollars suggests that the size of the stakes barely affects people's propensity to lie (Abeler et al. 2019). In *Golden Balls*, the stakes go far beyond the stakes typically used in laboratory experiments. For this larger range we do find some, albeit weak, evidence that the credibility of statements decreases when the amount of money involved increases: interactions of statement variables and the jackpot size are consistently negative and reach statistical significance in a few specifications.

The remaining subsections distinguish between the different types of implicit unconditional statements (Section 3.3) and examine whether contestants condition their cooperative choice on the statements of their opponent (Section 3.4).

3.3. Implicit Unconditional Statements

In the implicit unconditional category, there is a rich variety in the kind of statements that contestants make, whereas the other three contain relatively homogeneous sets. Contestants imply that they will split by (i) expressing

Figure 3. Cooperation Across Subcategories of Implicit Unconditional Statements



Note. This figure expands Figure 2(a) by splitting the category of implicit unconditional statements into six subcategories.

a preference to do so, (ii) referring to their character, (iii) referring to their past intention, (iv) expressing awareness that viewers will disapprove stealing, (v) indicating that half the jackpot is good enough or a lot of money, and (vi) urging the opponent to trust them.

Figure 3 displays the average cooperation rates across the six types, regardless of whether the contestant also made statements from any other (sub)category. There is considerable variation in cooperation, with character (75%) and viewer (76%) references being the most predictive of a split choice, followed by references to past intentions (67%), pleas for trust (60%), preference statements (59%), and money statements (48%).

Table 4 incorporates the subdivision of implicit unconditional statements in the regression models. Models 2 and 4 include the set of control variables. The six estimated marginal effect sizes rank similar as the average cooperation rates in Figure 3 and differ significantly, both when we use indicator variables and when we use proportions (both Wald $\chi^2(5) > 13.65$, both $p < 0.018$). Subdividing the implicit unconditional statements provides a notable improvement in the empirical fit. When indicator variables are used, the McFadden R^2 increases from 0.136 to 0.153 and the hit

percentage from 67.3 to 69.2. When proportions are used, the R^2 increases from 0.141 to 0.166 and the hit percentage from 67.3 to 68.7.

3.4. Opponent Statements

In the literature, there is considerable evidence that people are conditionally cooperative, in the sense that they have a preference for matching the cooperation of others (Fischbacher et al. 2001, Frey and Meier 2004). Contestants in *Golden Balls* appear to assume that their opponents are conditionally cooperative, because they generally try hard to signal their cooperative intention or hide their plan to steal. We have seen that some types of statements are clearly more predictive of cooperation than other statements. This raises the question whether contestants condition their decisions on the statements of their opponents.

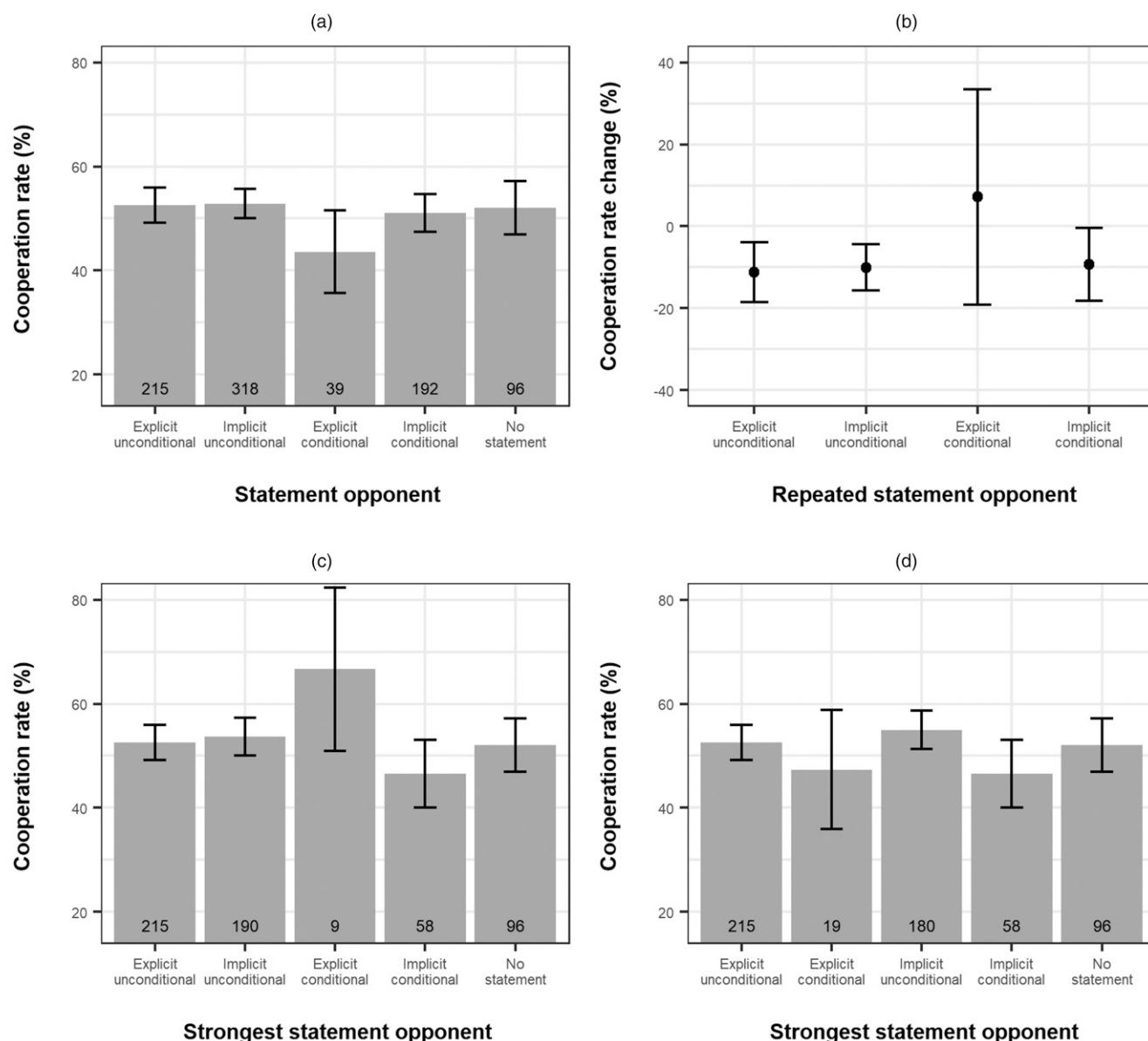
Figures 4(a) and 4(b) indicate that split or steal decisions do not depend on whether the opponent made a particular statement or not and that repetitions seem to be ignored. Figures 4(c) and 4(d) display the results for the opponent's strongest statement, using the two possible orderings of implicit unconditional and explicit conditional statements. Again, there seems

Table 4. Regression Results for Subcategories of Implicit Unconditional Statements

	Model 1	Model 2	Model 3	Model 4
	<i>One or more</i>	<i>One or more</i>	<i>Proportion</i>	<i>Proportion</i>
Explicit unconditional	0.202*** (0.044)	0.236*** (0.041)	0.390*** (0.058)	0.421*** (0.051)
Implicit unconditional				
Character	0.211** (0.082)	0.204*** (0.073)	0.745*** (0.187)	0.715*** (0.168)
Viewers	0.170* (0.098)	0.158* (0.084)	0.477** (0.215)	0.437** (0.191)
Past intention	0.123** (0.062)	0.134** (0.057)	0.321*** (0.101)	0.347*** (0.089)
Preference	0.102** (0.050)	0.124*** (0.048)	0.231*** (0.078)	0.254*** (0.071)
Trust	0.044 (0.080)	0.045 (0.074)	0.248* (0.139)	0.276** (0.132)
Money	−0.060 (0.044)	−0.030 (0.044)	0.018 (0.079)	0.056 (0.074)
Explicit conditional	0.061 (0.080)	0.042 (0.080)	0.267** (0.136)	0.283*** (0.108)
Implicit conditional	−0.022 (0.043)	−0.007 (0.042)	0.078 (0.063)	0.106* (0.057)
Controls	No	Yes	No	Yes
McFadden R^2	0.065	0.153	0.080	0.166
Number of clusters	284	284	284	284
Observations	568	568	568	568

Notes. The table reports the average marginal effects resulting from Logit regression analyses of contestants' decisions to split (1) or steal (0) the jackpot. The category of implicit unconditional statements is split into six subcategories. Models 1 and 2 use indicator variables measuring whether the contestant made at least one statement that belongs to the given (sub)category. Models 3 and 4 employ variables measuring the proportion of the contestant's statements that belong to the given (sub)category. Other definitions are as in Table 2.

* $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Figure 4. Cooperation and Opponent's Statements

Notes. Panel (a) displays the percentage of cooperators for all contestants whose opponent made at least one statement from a particular category and for all contestants whose opponent limited herself to empty talk. Panel (b) displays the differences in the percentages of cooperators between contestants whose opponent made at least two statements from a particular category and those whose opponent made only one. Panels (c) and (d) display the percentage of cooperators for all contestants whose opponent's strongest statement belongs to a particular category, with panel (c) assuming that implicit unconditional statements are stronger than explicit conditional statements, and panel (d) assuming the reverse. The number of contestants is at the bottom of each bar. Error bars depict standard errors around the mean.

to be no relation between contestants' propensity to cooperate and the strength of their opponent's statement.

Tables 5 and 6 present the regression results for opponent statements. Irrespective of whether we use indicator variables, proportions, or strongest-statement variables, a clear picture emerges: we are unable to reject the null hypothesis that the cooperative behavior of contestants is unaffected by the statements of their opponents. Virtually all explanatory power derives from contestants' own statements and the standard set of control variables.¹¹

In a similar vein, van den Assem et al. (2012) show that contestants do not condition their behavior on opponents' demographic characteristics, despite the predictive power of these. List (2006) and Oberholzer-Gee et al. (2010) report similar null results for the first season of a U.S. game show that resembles *Golden Balls*. Oberholzer-Gee et al. (2010), however, do observe conditioning on demographics in later seasons, suggesting that contestants learned to predict their opponent's behavior. In our data we find little evidence that conditioning arises as more episodes were

Table 5. Regression Results for Opponent's Statement Indicators and Proportions

	Model 1	Model 2	Model 3		Model 4	Model 5
	<i>One or more</i>	<i>One or more</i>	<i>One or more</i>	<i>Two or more</i>	<i>Proportion</i>	<i>Proportion</i>
Opponent						
Explicit unconditional	0.004 (0.046)	0.010 (0.042)	0.045 (0.046)	−0.099 (0.065)	0.011 (0.082)	−0.016 (0.073)
Implicit unconditional	0.008 (0.043)	−0.007 (0.040)	0.036 (0.043)	−0.089 (0.055)	0.020 (0.067)	−0.021 (0.060)
Explicit conditional	−0.094 (0.081)	−0.074 (0.073)	−0.075 (0.074)	0.042 (0.208)	−0.047 (0.163)	−0.052 (0.156)
Implicit conditional	−0.018 (0.045)	0.048 (0.043)	0.050 (0.045)	−0.030 (0.092)	−0.024 (0.075)	0.030 (0.072)
Contestant						
Explicit unconditional	— —	0.253*** (0.041)	0.251*** (0.047)	0.013 (0.062)	— —	0.428*** (0.050)
Implicit unconditional	— —	0.124*** (0.040)	0.085* (0.044)	0.089 (0.052)	— —	0.235*** (0.040)
Explicit conditional	— —	0.029 (0.080)	0.024 (0.084)	0.031 (0.179)	— —	0.277** (0.108)
Implicit conditional	— —	−0.018 (0.044)	−0.030 (0.046)	0.061 (0.086)	— —	0.106 (0.058)
Controls	No	Yes	Yes		No	Yes
McFadden R^2	0.002	0.139	0.150		0.001	0.142
Number of clusters	284	284	284		284	284
Observations	568	568	568		568	568

Notes. The table reports the average marginal effects resulting from Logit regression analyses of contestants' decisions to split (1) or steal (0) the jackpot. Models 1 and 2 use indicator variables measuring whether the contestant or her opponent made at least one statement that belongs to the given category. Model 3 adds variables indicating whether the contestant or her opponent made two or more statements belonging to the given category. Models 4 and 5 employ variables measuring the proportion of the contestant's statements or the proportion of her opponent's statements that belong to the given category. Other definitions are as in Table 2.

* $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

transmitted, neither on opponents' demographics nor on their statements (not tabulated). Belot et al. (2010) study a Dutch show and do find some evidence that an opponent's voluntary promise correlates with the contestant's likelihood of cooperation. Their result, however, is only marginally significant and only holds for a specific subset of contestants. All in all, high-stakes settings like *Golden Balls* thus provide little evidence for conditionally cooperative behavior.¹²

At first sight, the combination of predictable behavior and lack of conditioning seems to be at odds with the sizable literature on conditional cooperation. One possible explanation is that some contestants care about efficiency. In a weak variant of the Prisoner's Dilemma, such players will probably selfishly steal if they expect the other to split, but split if they are convinced that the other will steal because stealing would destroy the jackpot without any material gain. If sufficiently many contestants are in this category, their behavior would obfuscate the conditional cooperation of others in this game. Our video material, however, lends no support for this explanation. Episodes end with snippets of interviews in which contestants are asked to comment on their choice and the outcome of the game. If efficiency concerns

were real, we would expect at least some of the contestants who picked the split ball to claim that they expected the other to steal. Strikingly, no single contestant who chose split motivated her choice this way.

The likely explanation is that players with conditionally cooperative preferences fail to recognize the predictive power of their opponents' statements. Experimental studies in which subjects were shown video clips from *Golden Balls* or a similar show point in this direction, and suggest a more general inability to predict cooperative behavior: subjects perform at best only marginally better than chance when they try to predict contestants' decisions (Belot et al. 2012, Sylwester et al. 2012, Klein and Epley 2015).

4. Conclusions and Discussion

In this paper we depart from the conventional approach of analyzing the role of communication in cooperative choice in a binary way. The literature on free-form communication to date has mostly dichotomously classified statements as either "promises" or "empty." We hypothesize that lying aversion leads defectors to prefer statements that are malleable in terms of interpretation and allow them to deny—to themselves or to others—that they are

Table 6. Regression Results for the Opponent's Strongest Statement

Panel A: Implicit unconditional > explicit conditional		
	Model 1	Model 2
Opponent		
1. Explicit unconditional	0.005 (0.066)	−0.006 (0.060)
2. Implicit unconditional	0.016 (0.063)	−0.007 (0.058)
3. Explicit conditional	0.146 (0.167)	0.053 (0.186)
4. Implicit conditional	−0.055 (0.085)	−0.045 (0.078)
Contestant		
1. Explicit unconditional	—	0.396*** (0.055)
2. Implicit unconditional	—	0.211*** (0.056)
3. Explicit conditional	—	0.410*** (0.120)
4. Implicit conditional	—	0.097 (0.076)
Controls	No	Yes
McFadden R^2	0.002	0.145
Observations	284	284
Clusters	568	568
Panel B: Explicit conditional > implicit unconditional		
	Model 3	Model 4
Opponent		
1. Explicit unconditional	0.005 (0.066)	0.000 (0.060)
2. Explicit conditional	−0.047 (0.119)	−0.104 (0.106)
3. Implicit unconditional	0.029 (0.063)	0.010 (0.059)
4. Implicit conditional	−0.055 (0.085)	−0.040 (0.078)
Contestant		
1. Explicit unconditional	—	0.399*** (0.055)
2. Explicit conditional	—	0.292*** (0.105)
3. Implicit unconditional	—	0.215*** (0.056)
4. Implicit conditional	—	0.102 (0.076)
Controls	No	Yes
McFadden R^2	0.002	0.145
Observations	284	284
Clusters	568	568

Notes. The table reports the average marginal effects resulting from Logit regression analyses of contestants' decisions to split (1) or steal (0) the jackpot. The models use indicator variables for the strongest statement that the contestant or her opponent made. Models 1 and 2 in panel A assume that implicit unconditional statements are stronger than explicit conditional statements, and Models 3 and 4 in panel B assume the reverse. Other definitions are as in Table 2.

* $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

lying. Cooperators do not have such a preference, and, as a consequence, certain types of statements will be more indicative of cooperation than others.

We propose a typology that classifies statements according to two dimensions. This typology discriminates between statements that explicitly promise that the individual will cooperate and statements that only implicitly signal that she will do so, and between statements that do not carry any element of conditionality on the opponent's split or steal decision and statements that do carry such an element. We argue that explicit and unconditional statements are less malleable than implicit or conditional statements.

We have empirically examined the predictive power of the different types of statements using the British TV game show *Golden Balls*, where two contestants play a variant on the Prisoner's Dilemma with high stakes and free-form preplay communication. In line with the malleability hypothesis, our data show that explicit unconditional statements are associated with the highest likelihood of cooperation. Statements that either carry an element of conditionality or implicitness are generally associated with a moderate likelihood of cooperation. There is no clear reason to deem implicit unconditional statements more malleable than explicit conditional statements or vice versa, and in our data, there is no consistent evidence of such a difference. Last, contestants who make statements that are both conditional and implicit and contestants who do not make any statements related to their choice have the lowest rate of cooperation.

Within the category of implicit unconditional statements, we observe six different subcategories. These subcategories turn out to predict cooperation to varying degrees. The differences are not captured by our typology, but they do make intuitive sense. The cooperation rate is relatively high for contestants who referred to their character, to viewers' disapproval of defecting, or to their past intention. Character statements link the cooperative choice to the contestant's personality and values, which reveals awareness of and possibly amplifies the negative effects of defection on her self-concept and reputation. References to viewers specifically signal a contestant's reputation concerns. The high rate after past intention statements can be explained by people's desire to be and appear consistent (Cialdini 1984; Falk and Zimmermann 2017, 2018).

In spite of the predictive power, contestants do not condition their choice on their opponents' statements. This absence of conditioning implies that contestants either do not have a preference for matching the other's choice or cannot or do not make accurate assessments. The former explanation is not very likely given the evidence that contestants systematically reciprocate

opponents' earlier attempts to vote them off the show (van den Assem et al. 2012) and given contestants' frequent attempts to convince their opponents of their own cooperative intent. The latter explanation corresponds with experimental findings on people's ability to predict others' cooperative intentions. These findings are somewhat mixed, but the general picture is that forecasts are rather poor (Dawes et al. 1977, Frank et al. 1993, Brosig 2002, Belot et al. 2010, Sylwester et al. 2012, Klein and Epley 2015, He et al. 2017).

For the predictive power of malleability to hold, it is not required that people first decide whether they will cooperate and only then choose their statements accordingly. Malleability also predicts cooperation if people are undecided when they enter the conversation and base their decision on the statements they made. As a consequence of lying aversion, the option of stealing is costlier and thus less appealing for contestants whose statements were less malleable.

Our empirical results point out that our two-by-two typology provides a useful framework for categorizing the wide set of possible statements in free-form communication. We have applied it in the specific context of a variant on the Prisoner's Dilemma, but it can straightforwardly be used for a wider class of non-cooperative games.

A special feature of our field setting is that decisions are highly public, which may have amplified contestants' concerns about their reputation as a reliable person. At the same time, another special feature is that contestants are participating in what can be perceived as a "game," and lying in this game can potentially be seen as an inherent and unobjectionable element (Battigalli et al. 2013). In everyday life, most interactions are also public to some extent, and even in private interactions, people may worry about reputational repercussions of deceit because of gossip. Future experimental work could apply the typology to more anonymous conditions.

Acknowledgments

The authors thank Endemol Shine UK for providing them with information and recordings of *Golden Balls*. The paper has benefited from discussions with seminar participants at the Erasmus University of Rotterdam, the University of Michigan, the Berlin Social Science Center (WZB), the University of Amsterdam, the University of Reading, the University of Passau, and Ben-Gurion University of the Negev, and with participants of the NIBS 2014 Workshop, Coventry; the Risk, Uncertainty and Ambiguity 2014 Workshop, Ein Bokek; FUR 2014, Rotterdam; TIBER 2015, Tilburg; M-BEES 2015, Maastricht; ESA 2015, Heidelberg; the Behavioral Industrial Organization 2017 Workshop, Amsterdam; SPUDM 2017, Haifa; the FiBER 2018 Workshop, Haifa; the Minerva 2018 Spring Workshop, Technion; NCBE 2018, Odense; and the Cooperation: Interdisciplinary Methods 2018 Workshop, Lille.

Appendix. Definitions of Statement Categories and Examples

A.1. Explicit Unconditional

Definition: A statement is explicit unconditional (EU) if it (i) explicitly expresses that the contestant will choose split or not choose steal and (ii) carries no element of conditionality on the opponent's split or steal decision.

Examples: "I will split"; "I will not steal"; "I am going to split"; "I am going to pick up the split ball"; "There is no way I am going to pick the steal ball"; "My split ball is right there, and I am going to pick that ball up"; "I will not take it from you."

Nonexamples: "You know I am going to pick up the split ball"; "I told you before that I am going to split"; "You know I am going to pick up the split ball" (these statements are in the IU Plea for trust category).

A.2. Implicit Unconditional

Definition: A statement is implicit unconditional (IU) if it (i) does not explicitly express that the contestant will choose split or not choose steal but signals this through an indication of her preference (see IU preference), a reference to her character (see IU character), a reference to a past intention (see IU past intention), an expression of concern for viewers' or any other's judgment (see IU viewers), the opinion that half the jackpot is good or enough for her (see IU money), or because it urges her opponent to trust her (see IU plea for trust), and (ii) carries no element of conditionality on the opponent's split or steal decision.

A.2.1. IU Preference. *Definition:* IU preference statements signal that the contestant will choose split or not choose steal by expressing her current desire or intention to choose split, disapproval of stealing, or opinion that she should or feels obliged to choose split.

Examples: "I want to split with you"; "I am happy to split"; "I have got to split"; "I need to split"; "There is no need to be greedy and take everything home with me"; "I wouldn't steal" (this is quite close to IU character, but it is not the same as "I couldn't steal"; we interpret this statement as "I wouldn't want to steal"); "I would hate to steal and leave you empty-handed"; "I wouldn't feel good walking away with all the money"; "I think it would be stupid to steal"; "I think it's only fair to split"; "I don't want you to walk away with nothing"; "I would feel stupid stealing"; "Stealing is not worth losing your integrity over"; "I intend to split"; "I plan to split"; "My intention is to choose split"; "There is no point in being greedy"; "There is no point in stealing"; "It would be silly not to split"; "I think it is fair to share"; "I want to walk away with my integrity intact."

Nonexample: "I think I might split."

A.2.2. IU Character. *Definition:* IU character statements signal that the contestant will choose split or not choose steal through a reference to her character, nature, or general inability to steal.

Examples: "I am not the kind of person who steals"; "My conscience will not allow me to steal"; "I honestly do not think that I can steal"; "I couldn't steal"; "I am a splitter" (saying "I am the kind of person who splits"); "Stealing is not an option for me"; "I am not greedy."

Nonexamples: “I have been honest throughout the game” (description of own honesty in the past); “I am an honest person” (reference to honesty, which relates to what is being said and not to what is being done); “I cannot go home and look my eight year old in the eye and say that I stole it from another eight year old kid” (the “cannot steal” aspect is due to an other’s judgment; this statement is in the IU Viewers category); “I am a man of my word.”

A.2.3. IU Past Intention. *Definition:* IU past intention statements signal that the contestant will choose split or not choose steal through a reference to her past intention or desire to choose split, or to her intention or desire all along to choose split.

Examples: “I came here to split”; “I came here with the intention to split”; “I have always wanted to split”; “My game plan has always been to split”; “I am here to split” (refers to the contestant’s plan to choose split all along).

Nonexamples: “I came here to play an honest game” (honesty does not necessarily mean splitting); “I promised my class that if I came to this stage I would split the money” (although the statement refers to the past, the critical part relates to the judgment of others; this statement is in the IU viewers category).

A.2.4. IU Viewers. *Definition:* IU viewers statements signal that the contestant will choose split or not choose steal by expressing her awareness or worry that viewers will disapprove stealing or by referring to somebody else or some other people who want her to split.

Examples: “If I stole off you, this entire audience would lynch me”; “There is no way I am going to do you over on national TV”; “I cannot steal on TV”; “I cannot steal because my mom will be ashamed of me”; “I told my wife before coming on the show that I would split”; “I couldn’t face anybody if I stole”; “Everyone who knows me would be disgusted”; “I just think stealing would look so greedy you know” (it is implied that stealing would look greedy to others); “I promised my class that if I came to this stage I would split the money”; “My mom is watching. She told me to finish this with integrity”; “My son told me that I have got to split.”

Nonexamples: “My kids are watching”; “My entire family is watching” (mere descriptions of the situation, it is not sufficiently clear whether these people want the contestant to split or not).

A.2.5. IU Money. *Definition:* IU money statements signal that the contestant will choose split or not choose steal by expressing her feeling or opinion that half of the jackpot is a lot of money or good enough for her.

Examples: “I will be delighted to go home with half the jackpot”; “If I went home with 10,000 pounds, I would be ecstatic” (jackpot: £20,000); “10,000 pounds would make a difference” (jackpot: £20,000); “30,000 pounds is enough” (jackpot: £60,000); “I am looking to go home with 3,800 pounds” (interpreted as “I want to go home with 3,800 pounds”; jackpot: £7,600); “I want half the jackpot”; “I want you to have half the jackpot.”

Nonexample: “Half the jackpot is better than nothing” (doesn’t mean it is good enough).

A.2.6. IU Plea for Trust. *Definition:* IU Plea for trust statements signal that the contestant will choose split or not

choose steal by urging or pleading with the opponent to trust her or by indicating that the opponent can expect her to choose split.

Examples: “You can trust me”; “Trust me”; “I will not let you down”; “I will not rip you off” (indirectly saying that the opponent can have trust); “I won’t be greedy” (interpreted as “Trust me, I won’t be greedy”); “There is no way I am going to do you over” (indirectly saying that the opponent can have trust); “You know I am going to pick up the split ball” (interpreted as “you should know that I am going to pick up the split ball”); “I told you before that I am going to split” (indirectly indicating that the opponent has her word); “You have nothing to worry about”; “I guaranteed you in the last round that if you took me with I would split”; “You are going home with some money”; “I give you my word”; “You have my word.”

Nonexamples: “I don’t want to let you down” (interpreted as “I don’t want to steal”; this statement is in the IU Preference category); “I wouldn’t turn on you” (interpreted as “I would not want to turn on you”; this statement is in the IU Preference category); “I am a man of my word.”

A.3. Explicit Conditional

Definition: A statement is explicit conditional (EC) if it (i) explicitly expresses that the contestant(s) will choose split or not choose steal, and (ii) carries an element of conditionality on the opponent’s split or steal decision through an if-clause or a reference to the joint action.

Examples: “I will split if you split”; “I will split if you can assure me that you won’t let me down”; “We will split together”; “You and I will both split”; “I think you are going to do the right thing and share the money with me, and I am going to do the right thing and share the money with you”; “We are going to split it”; “I will steal if I think you are going to steal.”

A.4. Implicit Conditional

Definition: A statement is implicit conditional (IC) if it (i) does not explicitly express that the contestant(s) will choose split or not choose steal but signals it through an indication of a preference for both splitting, a reference to both contestants’ past intentions to split, a reference to viewers’ or any other’s judgment about the joint choice or outcome, the opinion that half the jackpot is good or enough for both, or because it urges for mutual trust, and (ii) carries an element of conditionality on the opponent’s split or steal decision through an if-clause or a reference to the joint action.

Examples: “I am willing to choose split if you will split”; “If you split with me, I am there for you all the way”; “We need to split the money”; “We should both go away with something”; “I would like us to split”; “We have to split”; “Let’s split”; “We said from the start that we would split”; “I think it should be a split” (split most likely refers to the joint outcome in this case); “I want you to feel the same as me when we go home”; “I think it is really awful if one of us stole”; “I want to split, and I would love for you to do the same for me”; “I want to split, but I want to be sure you’ll split with me”; “On national TV, whoever backed out is terrible”; “We need to trust each other now”; “I trust you and you trust me”; “I want you and me to have half the jackpot”; “We have got the money in our bank accounts” (conditional variant of “you are

going home with some money”); “We will go home happy”; “We might as well split”; “Let’s not be greedy”; “We will look silly if we steal” (interpreted as “We should split”); “It is just pure greed if we steal” (interpreted as “We should split”); “Half each.”

Nonexamples: “To get this far, and to both risk saying steal and going home with nothing, I would be absolutely gutted” (this is an opinion about both choosing steal, which is not the same as an opinion about both choosing split); “If we both steal, we both walk away with nothing” (description of the game); “If we split, we both walk away with 25,000” (description of the game); “I would like to end the show on a good note” (can mean anything); “525 pounds each” (describes half the jackpot, not the outcome).

Endnotes

¹ Even in experiments where subjects perform one-shot tasks under anonymous conditions, reputation concerns cannot be ruled out entirely, because subjects are normally well aware that their choices are recorded and scrutinized (Levitt and List 2007).

² Game show data have been widely used to study, for example, individual decision-making under risk (Gertner 1993, Metrick 1995, Beetsma and Schotman 2001, Post et al. 2008, Hartley et al. 2014), strategic decision-making (Bennett and Hickman 1993, Berk et al. 1996, Tenorio and Cason 2002), discrimination (Levitt 2004, Antonovics et al. 2005), cooperative behavior (List 2004, 2006; Belot et al. 2010; Oberholzer-Gee et al. 2010; van den Assem et al. 2012), and bargaining (van Dolder et al. 2015).

³ The reluctance to lie also depends on the type of deception. Lundquist et al. (2009) and Hilbig and Hessler (2013) show that people rather tell small lies than big lies. Gneezy (2005) and Erat and Gneezy (2012) find that people care about the consequences of a lie both for themselves and for others: the more someone can gain from lying, the more likely she is to lie, and the more a lie hurts the other, the less likely she is to lie. Spranca et al. (1991) demonstrate that people regard deception by omission as less immoral than deception through commission.

⁴ In the present paper, 215 contestants made one or more explicit unconditional statements. Out of those, 210 (98%) were assigned a value of one for the promise variable in the previous paper. The small difference can be attributed to previous coding errors. Out of the 305 contestants in the present paper who were assigned a value of one for the promise variable in the previous paper, 210 (69%) made one or more explicit unconditional statements. Two-thirds of the difference (63/95) can be ascribed to the different treatment of statements that have no meaning on their own, and the remainder to the present paper’s strict coding rules leaving little room for subjective interpretation.

⁵ Formally, malleability can also derive from conditionality on other unknowns, as long as the contestant can reasonably claim ignorance about whether the condition is satisfied until after she made her choice. In *Golden Balls*, however, conditional statements exclusively relate to the opponent’s split or steal decision.

⁶ Belot et al. (2010) find that game show contestants are indeed more likely to stick to a promise if it was made voluntarily than if it was elicited by a question from the host.

⁷ Video clips of this and many other episodes are widely available on the internet, for example, through YouTube.

⁸ In most cases, the conditionality derives from a reference to the joint action. Only nine explicit conditional statements and eight implicit conditional statements are through an if-clause.

⁹ For comparison: the intercept-only model has a hit percentage of 52.5.

¹⁰ Both Dreber and Johannesson (2008) and Erat and Gneezy (2012) find that men are more willing to lie in sender-receiver games than women, whereas Childs (2012) and Gylfason et al. (2013) find no significant difference. Abeler et al. (2019) and Capraro (2018) perform a meta-analysis of lying experiments and conclude that men lie more than women.

¹¹ Distinguishing between the six different types of implicit unconditional statements made by the opponent adds no significant explanatory power (not tabulated).

¹² We have in addition investigated the behavior of contestants for whom there is a particular reason to believe that they are conditionally cooperative. Episodes contain fragments of private interviews that individual contestants had with the producer, in which they announce their game plan. We found no evidence that contestants who indicated that their decision depends on their opponent cooperate more when their opponent’s demographics or statements are more indicative of cooperation. Similarly, we found little evidence of conditioning among contestants who made (explicit or implicit) conditional statements. The only exception is that these contestants are more likely to cooperate when their opponent made an explicit unconditional statement. Note, however, that this evidence for conditional cooperation also aligns with our malleability hypothesis, because such an unmalleable statement by the opponent makes it harder for a contestant who made a conditional statement to steal and then deny that she has lied. Last, we could not find any evidence that contestants are matching their opponents’ actual choice, neither for the two types of contestants who indicated to be conditionally cooperative nor for the full sample.

References

- Abeler J, Becker A, Falk A (2014) Representative evidence on lying costs. *J. Public Econom.* 113:96–104.
- Abeler J, Nosenzo D, Raymond C (2019) Preferences for truth-telling. *Econometrica*. Forthcoming.
- Antonovics K, Arcidiacono P, Walsh R (2005) Games and discrimination: Lessons from the weakest link. *J. Human Resources* 40(4): 918–947.
- Battigalli P, Charness G, Dufwenberg M (2013) Deception: The role of guilt. *J. Econom. Behav. Organ.* 93:227–232.
- Beetsma RMWJ, Schotman PC (2001) Measuring risk attitudes in a natural experiment: Data from the television game show Lingo. *Econom. J.* 111(474):821–848.
- Belot M, Bhaskar V, van de Ven J (2010) Promises and cooperation: Evidence from a TV game show. *J. Econom. Behav. Organ.* 73(3): 396–405.
- Belot M, Bhaskar V, van de Ven J (2012) Can observers predict trustworthiness? *Rev. Econom. Statist.* 94(1):246–259.
- Bennett RW, Hickman KA (1993) Rationality and ‘the price is right’. *J. Econom. Behav. Organ.* 21(1):99–105.
- Berk JB, Hughson E, Vandezande K (1996) The Price Is Right, but are the bids? An investigation of rational decision theory. *Amer. Econom. Rev.* 86(4):954–970.
- Brosig J (2002) Identifying cooperative behavior: Some experimental results in a Prisoner’s Dilemma game. *J. Econom. Behav. Organ.* 47(3):275–290.
- Cappelen AW, Sorensen EO, Tungodden B (2013) When do we lie? *J. Econom. Behav. Organ.* 93:258–265.
- Capraro V (2018) Gender differences in lying in sender-receiver games: A meta-analysis. *Judgment Decision Making* 13(4):345–355.
- Charness G, Dufwenberg M (2006) Promises and partnership. *Econometrica* 74(6):1579–1601.
- Charness G, Dufwenberg M (2010) Bare promises: An experiment. *Econom. Lett.* 107(2):281–283.
- Childs J (2012) Gender differences in lying. *Econom. Lett.* 114(2): 147–149.

- Cialdini RB (1984) *Influence: The Psychology of Persuasion* (Harper Collins, New York).
- Dawes RM, McTavish J, Shaklee H (1977) Behavior, communication, and assumptions about other people's behavior in a commons dilemma situation. *J. Personality Soc. Psych.* 35(1):1–11.
- DePaulo BM, Lindsay JJ, Malone BE, Muhlenbruck L, Charlton K, Cooper H (2003) Cues to deception. *Psych. Bull.* 129(1):74–118.
- Dreher A, Johannesson M (2008) Gender differences in deception. *Econom. Lett.* 99(1):197–199.
- Ederer F, Stremitz A (2017) Promises and expectations. *Games Econom. Behav.* 106:161–178.
- Ekman P (2001) *Telling lies: Clues to Deceit in the Marketplace, Politics, and Marriage* (W.W. Norton, New York).
- Ellingsen T, Johannesson M (2004) Promises, threats and fairness. *Econom. J.* 114(495):397–420.
- Erat S, Gneezy U (2012) White lies. *Management Sci.* 58(4):723–733.
- Falk A, Zimmermann F (2017) Consistency as a signal of skills. *Management Sci.* 63(7):2197–2210.
- Falk A, Zimmermann F (2018) Information processing and commitment. *Econom. J.* 128(613):1983–2002.
- Farrell J, Rabin M (1996) Cheap talk. *J. Econom. Perspect.* 10(3):103–118.
- Fischbacher U, Gächter S, Fehr E (2001) Are people conditionally cooperative? Evidence from a public goods experiment. *Econom. Lett.* 71(3):397–404.
- Frank RH, Gilovich T, Regan DT (1993) The evolution of one-shot cooperation: An experiment. *Ethology Sociobiol.* 14(4):247–256.
- Frey BS, Meier S (2004) Social comparisons and pro-social behavior: Testing 'conditional cooperation' in a field experiment. *Amer. Econom. Rev.* 94(5):1717–1722.
- Gertner R (1993) Game shows and economic behavior: Risk-taking on Card Sharks. *Quart. J. Econom.* 108(2):507–521.
- Gneezy U (2005) Deception: The role of consequences. *Amer. Econom. Rev.* 95(1):384–394.
- Gylfason HF, Arnardottir AA, Kristinsson K (2013) More on gender differences in lying. *Econom. Lett.* 119(1):94–96.
- Hartley R, Lanot G, Walker I (2014) Who really wants to be a millionaire? Estimates of risk aversion from gameshow data. *J. Appl. Econometrics* 29(6):861–879.
- He S, Offerman T, van de Ven J (2017) The sources of the communication gap. *Management Sci.* 63(9):2832–2846.
- Hilbig BE, Hessler CM (2013) What lies beneath: How the distance between truth and lie drives dishonesty. *J. Experiment. Soc. Psych.* 49(2):263–266.
- Khalmetski K, Rockenbach B, Werner P (2017) Evasive lying in strategic communication. *J. Public Econom.* 156:59–72.
- Klein N, Epley N (2015) Group discussion improves lie detection. *Proc. Natl. Acad. Sci. USA* 112(24):7460–7465.
- Levitt SD (2004) Testing theories of discrimination: Evidence from weakest link. *J. Law Econom.* 47(2):431–453.
- Levitt SD, List JA (2007) What do laboratory experiments measuring social preferences reveal about the real world? *J. Econom. Perspect.* 21(2):153–174.
- List JA (2004) Young, selfish and male: Field evidence of social preferences. *Econom. J.* 114(492):121–149.
- List JA (2006) Friend or foe? A natural experiment of the Prisoner's Dilemma. *Rev. Econom. Statist.* 88(3):463–471.
- Lundquist T, Ellingsen T, Gribbe E, Johannesson M (2009) The aversion to lying. *J. Econom. Behav. Organ.* 70:81–92.
- Mazar N, Amir O, Ariely D (2008) The dishonesty of honest people: A theory of self-concept maintenance. *J. Marketing Res.* 45(6):633–644.
- Metrick A (1995) A natural experiment in Jeopardy! *Amer. Econom. Rev.* 85(1):240–253.
- Oberholzer-Gee F, Waldfogel J, White MW (2010) Friend or foe? Cooperation and learning in high-stakes games. *Rev. Econom. Statist.* 92(1):179–187.
- Pittarello A, Leib M, Gordon-Hecker T, Shalvi S (2015) Justifications shape ethical blind spots. *Psych. Sci.* 26(6):794–804.
- Post T, van den Assem MJ, Baltussen G, Thaler RH (2008) Deal or No Deal? Decision making under risk in a large-payoff game show. *Amer. Econom. Rev.* 98(1):38–71.
- Rapoport A (1988) Experiments with N-person social traps I: Prisoner's Dilemma, Weak Prisoner's Dilemma, Volunteer's Dilemma, and Largest Number. *J. Conflict Resolution* 32(3):457–472.
- Schweitzer ME, Hsee CK (2002) Stretching the truth: Elastic justification and motivated communication of uncertain information. *J. Risk Uncertainty* 25(2):185–201.
- Serra-Garcia M, van Damme E, Potters J (2011) Hiding an inconvenient truth: Lies and vagueness. *Games Econom. Behav.* 73(1):244–261.
- Shalvi S, Dana J, Handgraaf MJJ, De Dreu CKW (2011) Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior. *Organ. Behav. Human Decision Processes* 115(2):181–190.
- Sporer SL, Schwandt B (2007) Moderators of nonverbal indicators of deception: A meta-analytic synthesis. *Psych. Public Policy Law* 13(1):1–34.
- Spranca M, Minsk E, Baron J (1991) Omission and commission in judgment and choice. *J. Experiment. Soc. Psych.* 27(1):76–105.
- Sylwester K, Lyons M, Buchanan C, Nettle D, Roberts G (2012) The role of theory of mind in assessing cooperative intentions. *Personality Individual Differences* 52(2):113–117.
- Tenorio R, Cason TN (2002) To spin or not to spin? Natural and laboratory experiments from The Price Is Right. *Econom. J.* 112(476):170–195.
- Vanberg C (2008) Why do people keep their promises? An experimental test of two explanations. *Econometrica* 76(6):1467–1480.
- van den Assem MJ, van Dolder D, Thaler RH (2012) Split or steal? Cooperative behavior when the stakes are large. *Management Sci.* 58(1):2–20.
- van Dolder D, van den Assem MJ, Camerer CF, Thaler RH (2015) Standing united or falling divided? High stakes bargaining in a TV game show. *Amer. Econom. Rev.* 105(5):402–407.
- Vrij A (2008) *Detecting Lies and Deceit: Pitfalls and Opportunities*, 2nd ed. (John Wiley & Sons, Chichester, UK).
- Wooldridge JM (2003) Cluster-sample methods in applied econometrics. *Amer. Econom. Rev.* 93(2):133–138.
- Zuckerman M, DePaulo BM, Rosenthal R (1981) Verbal and non-verbal communication of deception. Berkowitz L, ed. *Advances in Experimental Social Psychology*, vol. 14 (Elsevier, New York), 1–59.