Lies in Disguise – A Theoretical Analysis of Cheating*

Martin Dufwenberg, Jr. & Martin Dufwenberg, Sr.

November 6, 2016

Abstract: We perform a (psychological) game-theoretic analysis of cheating in the setting proposed by Fischbacher & Föllmi-Heusi (2013). The key assumption, which we refer to as *perceived cheating aversion*, is that the decision maker derives disutility in proportion to the amount in which he is perceived to cheat. A particular equilibrium captures the stylized facts from many experiments (in particular the co-presence of selfish, honest, and partial-lie choices) well.

Keywords: cheating, lying, costs, audience, perceived cheating aversion

***JEL* codes:** C72, D03, D82, D83, H26

.

*MDJr: School of Life Sciences, Arizona State University; martindufwenberg@yahoo.com. MDSr: Department of Economics, University of Arizona and University of Gothenburg; martind@eller.arizona.edu. We started this paper (1st draft 4/27/16) when one of us was Uri Gneezy's RA on the project "Why Don't People Lie More?" We thank Uri for early encouragement, the UCSD Rady School for hospitality, and Johannes Abeler, Navin Kartik, Daniele Nosenzo, and Collin Raymond for their comments.

1 Introduction

Situations are aplenty where cheating may be materially lucrative. Examples involve tax evasion, embezzlement of foreign aid, or scientific misconduct to increase likelihood of a good publication. However, aspects of social rewards or personal integrity may mitigate opportunistic behavior.[1]

Researchers grapple with how to disentangle the impact of morale or honesty from that of formal sanctions (e.g. fines). In this connection lab experiments, which can control, and often then rule out, formal sanctions may be helpful. A recent literature, pioneered by Mazar, Amir & Ariely (2008) and Fischbacher & Föllmi-Heusi (F&FH) (2013, but written much earlier), that uses such methods has developed rapidly.[2] Most studies build on F&FH's design: Subjects roll a die (or flip a coin), self-report the outcome, and get paid based on the report. Although it is impossible to detect lying on an individual level, cheating behavior across the sample population can be quantified as the experimenters knows the underlying distribution. F&FH report that 20% of people lie to the fullest extent, 39% choose as if honest, and a sizeable proportion cheat a bit. Other scholars report similar findings. Various explanations have been proposed – F&FH themselves consider lying aversion, caring about the credibility of the lie, and the notion of self-concept maintenance first proposed by Mazar et al.

We propose a new theoretical model and explore cheating in settings close to that of F&FH (and to signal that link, our paper's title parallels

[1] For related commentary, see e.g. Luttmer & Singhal (2014) on tax morale and Olken (2015, p. 76) on honesty in research.

[2] See Abeler, Nosenzo & Raymond's (2016) survey, meta-study (based on 72 studies), summary of explanations, and new experimental tests. A sample of studies that use F&FH's paradigm include Abe & Greene (2014), Abeler et al. (2014, 2016), Arbel et al. (2014), Bucciol & Piovesan (2011), Cohn et al (2014, 2015), Conrads et al. (2013), Conrads & Lotz (2015), Dai et al. (2016), Dieckmann et al. (2015), Diekmann et al. (2015), Fosgaard et al. (2013), Gächter & Schulz (2016), Gneezy et al. (2016), Houser et al. (2012, 2016), Kajackaite & Gneezy (2016), Kocher et al. (2016), Kroher & Wolbring (2015), Muehlheusser et al. (2015), Pascual-Ezama et al. (2014), Piff et al. (2012; "game of chance"), Shalvi et al. (2010, 2011, 2012), Utikal & Fischbacher (2013).

theirs). We imagine a scenario where an audience (e.g. a tax authority, a granting agency, an editor, or an experimenter) observes the report a decision maker (henceforth, "DM") issues regarding a random outcome that only DM observes. By issuing a false report DM can mislead the audience with impunity and gains financially. Our central assumption is that DM feels bad to the extent that the audience believes he is cheating, a sentiment we refer to as *perceived cheating aversion*. Therefore reporting the outcome that brings DM the highest profit may not necessarily yield the greatest utility. DM will cheat and lie if he can do so credibly, but the audience is smart and draws inferences based on an understanding of his incentives, which may make DM less inclined to cheat.

Perceived cheating aversion makes DM's utility belief-dependent in the sense of psychological game theory (Geanakoplos, Pearce & Stacchetti 1989; Battigalli & Dufwenberg 2009). As it turns out, a particular sequential equilibrium (Example 4) allows us to capture the central tendencies of F&FH data remarkably well (in particular the co-presence of selfish, honest, and partial-lie choices). There are caveats and twists to that conclusion, including considerations regarding how our theory relates to Mazar et al's notion of self-concept maintenance. We postpone a discussion until we have introduced our assumptions formally and derived our results.

Gneezy, Kajackaite & Sobel (2016), Khalmetski & Sliwka (2016), as well as Abeler, Nosenzo & Raymond (2016, see e.g. section 2.3) perform exercises which are closely related in that they too examine forms of belief-dependent utility: Their DM gets disutility proportional to the probability that others assign that DM lies, while our DM gets disutility proportional to the amount in which he is perceived to cheat.[3] (Unlike us, GK&S and K&S, and in some

[3] Also related are the models of Bernheim (1994) on "status/conformity," Dufwenberg & Lundholm (2001) on "social respect," Battigalli & Dufwenberg (2007) on "guilt from blame," Ellingsen & Johannesson (2008) on "pride & prejudice," Tadelis (2008) on "shame," and Andreoni & Bernheim (2009) on "social image." Game forms and utilities differ greatly, but in all cases a player cares about others' inferences/opinions regarding own choice or type or motive, rendering these models psychological games. See Battigalli & Dufwenberg's (2009) discussion of "R2" on p. 7 for some relevant related commentary.

sections AN&R, also let utilities reflect direct costs of lying.)

Section 2 presents the game forms we consider, and some background results. Section 3 introduces our main psychological assumption, and explores equilibrium behavior. Section 4 relates our findings to data and ideas in the preceding literature.

2 Preliminaries

Game Form Nature randomly draws $x \in \{0, ..., n\}$, $n \geq 1$, from a uniform distribution. A decision maker (DM) observes and is asked to report x, but the report is non-verifiable so DM may choose any $y \in \{0, ..., n\}$ and is then paid $\$T \cdot y$, where $T > 0$ reflects the stakes. The chosen y (but not x) is observed by an audience (e.g. a neighbor, tax authority, or experimenter). A (behavioral) strategy for DM is a function $s : \{0, ..., n\} \to \Delta\{0, ..., n\}$, so $s(x)(y)$ denotes the probability s assigns to y after DM observes x. If $s(x)(y) = 1$, we sometimes write $s(x) = y$. S denotes DM's set of strategies.

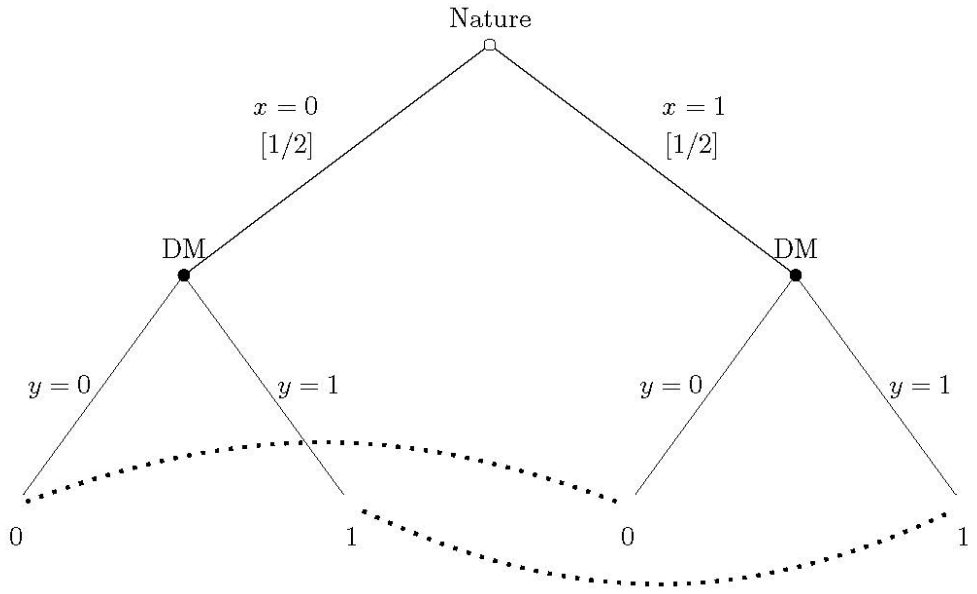
This set-up (essentially) follows F&F-H who consider the (die-roll) case where $n = 5$.[4] Others studied the (coin-flip) case where $n = 1$;[5] the associated game-tree is illustrated in Figure 1:

.

[4] In F&FH's setup the DM is paid $T \cdot y$ if $y \in \{1, 2, 3, 4, 5\}$ and paid 0 if $y = 6$.

[5] Bucciol & Piovesan (2011) may have been first to do this.

Figure 1, n=1 (coin-flip)



The numbers at the end-nodes are DM's monetary payoffs, not his utilities which are yet to be defined. The information-sets across end-nodes reflects the audience's end-of-play information. The audience has no choice and our analysis will not depend on its payoff, so we do not specify that.

Direct Cheating Costs These reflect that DM feels bad if he chooses $y > x$. This sentiment isn't our main focus but itforms a useful yardstick for comparison. It can be modeled such that DM's utility at end node (x, y) (= path x-then-y) depends negatively on the amount cheated, $T \cdot [y-x]^+$, where $[y-x]^+ = \max\{y-x, 0\}$. If DM's utility equals $T \cdot y - \theta \cdot T \cdot [y-x]^+$ we get the special case of linear direct cheating costs, where $\theta \geq 0$ is a parameter measuring DM's sensitivity/aversion to cheating. DM's optimal choice does not depend on T: If $\theta < 1$ he chooses $y = n$ regardless of x. If $\theta = 1$ he may choose any element of $\Delta\{x, ..., n\}$. If $\theta > 1$ he chooses $y = x$.

To escape this all-or-nothing (selfish-or-honest) implication, one may consider convex costs which may yield partial lies of specific step-length, that furthermore depends on T. To illustrate, suppose utility equals $T \cdot y - \theta \cdot (T \cdot [y-x]^+)^2$. Verify that if $\theta = \frac{2}{9}$ and $T = 1$ then DM chooses $y = \max\{x+2, n\}$, while if $\theta = \frac{2}{9}$ and $T = 2$ he chooses $y = \max\{x+1, n\}$.

Direct cheating costs are closely comparable to lying costs, which could be linear or convex in the size of the lie (here $|y-x|$); cf. e.g. Kartik (2009). The interpretation is slightly different (the dimension of cheating is money rather than distance in some message space, and cheating entails no cost of downward lies), but the behavioral conclusions are obviously analogous.

3 Playing to the Audience

We now consider our central assumption: *DM feels bad to the extent that the audience believes that he is cheating.* We explore this while abstracting away from direct cheating costs. It may be plausible to combine direct cheating costs with dislike that others believe one cheats, but in order to highlight the insights that come from the latter sentiment we disregards the former except as a yardstick for comparison.

3.1 Beliefs, Utility, Solution Concept

At end-node (x, y), DM's actual amount of cheating equals $T \cdot [y - x]^+$. The audience does not observe x, but draws inferences conditional on y. Let $p(x'|y) \in [0, 1]$ be the probability it assigns to $x = x'$ given y, with $\sum_{x'} p(x'|y) = 1$, so the audience's expectation of DM's cheating equals

$$\sum_{x'} (p(x'|y) \cdot T \cdot [y - x']^+), \tag{1}$$

of which DM abhors high values. Namely, DM's utility at (x, y) equals

$$T \cdot y - \theta \cdot \sum_{x'} (p(x'|y) \cdot T \cdot [y - x']^+), \tag{2}$$

where again $\theta \geq 0$ measures sensitivity. This utility depends on the audience's beliefs, via $p(x'|y)$, so we have a psychological game in the sense of Geanakoplos et al. and Battigalli & Dufwenberg (2009).[6] Note that the utility is independent of x. DM cares about his image, not about cheating per se. We say that (2) captures *perceived cheating aversion.*

We focus on equilibria where DM uses a strategy that maximizes his utility while the audience has correct initial beliefs and uses Bayes' rule to draw inferences regarding DM's cheating:

Definition: $s \in S$ is a **sequential equilibrium** (**SE**) if for $x, x', y \in \{0, ..., n\}$ we have $p(x'|y) \in [0,1]$ and $\sum_{x'} p(x'|y) = 1$ and the following two conditions holds:

(i) $s(x)(y) > 0 \Rightarrow y \in \arg\max_z (T \cdot z - \theta \cdot \sum_{x'} (p(x'|z) \cdot T \cdot [z - x']^+)$,

(ii) $\sum_x s(x)(y) > 0 \Rightarrow p(x'|y) = \frac{s(x')(y)}{\sum_x s(x)(y)}$.

Condition (i) says, wrt (2), that s maximizes DM's utility given the audience's inferences. Condition (ii) says that if y is a choice DM may make when using s (in the sense that $\sum_x s(x)(y) > 0$) then $p(x'|y)$ is calculated using Bayes' rule based on correct initial audience beliefs. If y is a choice DM will not make when using s (so that $\sum_x s(x)(y) = 0$) then no restrictions apply to $p(x'|y)$ except that $p(x'|y) \in [0,1]$ and $\sum_{x'} p(x'|y) = 1$. The SE-terminology is justified in that predictions coincide with Kreps & Wilson's (1982) classic notion if $\theta = 0$, and if $\theta \geq 0$ with Battigalli & Dufwenberg's (2009) extension of SE to psychological games.[7]

[6] We need B&D's framework as GP&S would not allow DM's utility to depend on another's beliefs or on an updated belief; $p(x'|y)$ has both features, being the audience's updated belief. The models of Gneezy et al. and Khalmetski & Sliwka similarly fit B&D's framework, although their functional forms differ from ours. Their DM gets disutility proportional to the probability the audience attaches to a lie ($\sum_{x' \neq y} p(x'|y)$), unmodulated by the amount $[y - x']^+$ in which he is perceived to cheat (so unlike (2)).

[7] B&D specify conditional beliefs about beliefs of all orders while, to reduce notation, we detail only the beliefs that matter to DM's utility ($p(x'|y)$) and leave other beliefs implicit including that DM correctly anticipates $p(x'|y)$.

3.2 Equilibrium Analysis

Our narrative comprises a series of arguably pedagogically linked observations and examples. We first establish two basic properties:

Observation 1: (a) At least one SE always exist. (b) The set of SEs does not depend on T (across games with the same n and θ).

Proof: Part (a) follows as B&D prove existence of SE for a large class of psychological games to which ours belong (and below we actually construct SE for all parameter constellations). Part (b) follows by inspection of condition (i) of the Definition. ■

The next two results rule out honesty and, for high enough θ, selfishness:

Observation 2: $s \in S$ def. by $s(x) = x$ for all x is not a SE for any θ.

Proof: By (ii) of the Definition, if $x' \neq y$ then $p(x'|y) = \frac{0}{1+n\cdot 0} = 0$. In this case (1) equals 0, so DM's utility from y, given by (2), reduces to $T \cdot y - 0 = T \cdot y$, which is maximized if DM deviates to choose $y = n$ regardless of x. ■

Intuitively, if honesty were an equilibrium, then the audience would believe that $x = y$, so that cheating at $y = n > x$ would go undetected.

Observation 3: $s \in S$ defined by $s(x) = n$ for all x is a SE iff $\theta \leq 2$.

Proof: DM's utility cannot be negative, as DM could always get 0 by choosing $y = 0$ (no cheating is possible then). So, using (2), we get

$$T \cdot n - \theta \cdot \sum_{x} (p(x|n) \cdot T \cdot [n - x]^{+}) \geq 0. \tag{3}$$

The audience has correct initial beliefs, and by (ii) of the Definition and since Nature's choice of x is drawn from a uniform distribution we get that $p(x'|n) = \frac{1}{n+1}$ for all $x' \in X$, implying that the lhs of (3) equals $Tn - \theta \cdot T \cdot \frac{n}{2}$, so (3) cannot hold if $\theta > 2$. On the other hand, on setting $p(0|y) = 1$ for all $y < n$, one infers (using (3)) that $s(x) = n$ for all x is a SE for any $\theta \leq 2$. ■

Intuitively, pooling on $y = n$ implies strong perceived cheating; if $\theta > 2$ DM would escape that impression by deviating to $y = 0$.

Observations 1, 2, and 3 combine to imply that for high enough θ SE exist but must exhibit partial lies, in stark contrast with the all-or-nothing conclusions associated with direct costs of cheating of section 2. To provide a feel for the bouquet of patterns that SE admits, and some relevant constraints, we now present a series of examples.

When experiments are run all choices $y \in \{0, ..., n\}$ tend to occur. We make ourselves no illusion that such data necessarily reflects equilibrium play. Nevertheless, it is intriguing to ask whether SE exist such that any y occurs with positive probability in the sense that $\sum_x s(x)(y) > 0$. We say that any such s has **full-support-on-**y. The associated utility following any choice y must equal 0. This is because 0 is what that DM gets if he chooses $y = 0$ (as noted in the proof of Observation 3) and in equilibrium all choices of y must give the same utility (otherwise there would be a profitable deviation). We say that s then has the 0**-utility-property**.

We start with two examples which, while highly special, exhibit intriguing cheating patterns and which pave way for instructive insights on the role of n and θ in shaping which SE are feasible:

Example 1: [**"rotten-0"**] Consider $s \in S$ such that:

$s(0)(y) > 0$ for all y,

$s(x) = x$ for all $x \neq 0$,.

That is, cheating occurs only when $x = 0$, but then involve all values of y. For s to be a SE, by the 0-utility-property and using (2), we get

$T \cdot y - \theta \cdot p(0|y) \cdot T \cdot y = 0$. For $y > 0$ we get $p(0|y) = \frac{1}{\theta}$ and since $p(0|y) = \frac{s(0)(y)}{s(0)(y)+1}$ we get $\frac{1}{\theta} = \frac{s(0)(y)}{s(0)(y)+1}$ or $s(0)(y) = \frac{1}{\theta-1}$. Furthermore, we have that $s(0)(y) < \frac{1}{n}$ (as otherwise we would get $\sum_{1 \leq y \leq n} s(0)(y) = 1$ so that $s(0)(0) > 0$ could not hold), implying $\frac{1}{\theta-1} < \frac{1}{n}$, or $\theta > n + 1$. ▲

Example 2: ["one-upmanship"] Consider $s \in S$ such that:

for all $x < n$, $s(x)(y) > 0$ iff $y \in \{x, x+1\}$,

$s(n) = n$.

That is, lies occur with positive probability for all $x < n$, but then involve only the smallest exaggeration with one-unit cheating. The SE can be constructed recursively. By the 0-utility-property, and using (2), we get $T \cdot y - \theta \cdot p(y-1|y) \cdot T \cdot 1 = 0$. For $y > 0$ we get $p(y-1|y) = \frac{y}{\theta}$ and by (ii) of the Definition we have $p(y-1|y) = \frac{s(y-1)(y)}{s(y-1)(y)+s(y)(y)}$, so that

$$\frac{y}{\theta} = \frac{s(y-1)(y)}{s(y-1)(y) + s(y)(y)}, \tag{4}$$

and hence $s(y-1)(y) = \frac{y \cdot s(y)(y)}{\theta - y}$. For $y = n$, since $s(n)(n) = 1$, this becomes $s(n-1)(n) = \frac{n}{\theta-n}$. It must also hold that $s(n-1)(n) < 1$, as otherwise the property that $s(n-1)(n-1) > 0$ would be violated. Combining these constraints on $s(n-1)(n)$ implies that $\frac{n}{\theta-n} < 1$, or $\theta > 2n$. It is straightforward to now verify, making repeated use of (4), that $s(y-1)(y)$ is well-defined for all $y \in \{1, ..., n\}$, and that $0 < s(y-1)(y) < s(n-1)(n)$.[8] ▲

The behaviors exhibited in Examples 1 and 2 are striking in terms of the idiosyncratic cheating patterns. However, a large n undermines the existence of such an SE, for any given θ. The respective necessary conditions, $\theta > n+1$ and $\theta > 2n$, simply cannot hold if n is large enough, for a given θ. To get a better feel for why this is so, note that in Example 1, to satisfy the 0-utility-property, all cheating is done when $x = 0$ occurs. With $s(x)(x) = 1$

[8] Recursively consider $y = n-1, n-2, ..., 1$. To reach the insight that $s(y-1)(y) < s(n-1)(n)$, it is key to note that the numerator of the lhs of (4) is increasing in y, and that in the denominator of the rhs of (4) it holds that $s(y)(y) < s(n)(n) = 1$.

for all $x > 0$, to achieve $p(0|x) = \frac{1}{\theta}$, the probability $s(0)(x)$ must reach a critical strictly positive level for all $x > 0$, independently of what n is. This becomes impossible as n grows because there is only a unit of probability to play with. In Example 2, to achieve the 0-utility-property as regards $y = n$, all the cheating is done when $x = n - 1$. Since the cheating going on when $x = n - 1$ and $y = n$ is small (namely, $y - x = n - (n - 1) = 1$), and since to achieve the 0-utility-property the drop in utility must overcome that the material reward grows with n. For a given θ, $s(n-1)(n)$ must be proportional to n, which is impossible as there is only a unit of probability to play with.

The following result generalizes those insights:

Observation 4: Fix $\theta > 0$ and $k \in \mathbb{N}_0$. (a) Consider any full-support strategy $s \in S$ such that $s(x) = x$ for all $x > k$. If n is large enough then s is not a SE. (b) Consider any full-support strategy $s \in S$ such that $s(n) = n$ while for all $x < n$ we have $s(x)(y) > 0$ only if $y \in \{x, x + k\}$. If n is large enough then s is not a SE. ■

We omit the proof; the essence mirrors the intuition highlighted via Examples 1 and 2. The upshot: neither cheating-only-at-a-low-x's nor lowball-cheating-for-all-x's can be sustained as SE, for large enough n. And the flip-side is (as with Examples 1 and 2) that if θ is too low then s as described in Observation 4 is not an SE. So to have SE for low values of θ or high values of n, more drastic patterns of lies or cheating are called for. The following example describes a class of such full-support SE which turn out to be viable for any n. Moreover, θ may take values so low as to allow identification of its infimum to allow non-selfish play as an SE. The behavioral patterns described involve the perhaps unexpected feature that lies are systematic but mainly directed *downwards*, i.e. DM often chooses $y < x$.

Example 3: [**"crowded-floor"**] Consider $s \in S$ such that:

$s(0)(y) > 0$ for all y,

for all $x \neq 0$, $s(x)(0) = 1 - \varepsilon$ and $s(x)(x) = \varepsilon \in (0,1)$.

Cheating occurs only when $x = 0$, and involves all values of y. However, (non-cheating) lies occur also when $x > 0$ as with probability $1 - \varepsilon$ DM then *under-reports* by choosing $y = 0$. For s to be a SE, by the 0-utility-property and using (2), we get $T \cdot y - \theta \cdot p(0|y) \cdot T \cdot y = 0$. For $y > 0$ we get $p(0|y) = \frac{1}{\theta}$ and since $p(0|y) = \frac{s(0)(y)}{s(0)(y)+\varepsilon}$ we get $\frac{1}{\theta} = \frac{s(0)(y)}{s(0)(y)+\varepsilon}$ or $s(0)(y) = \frac{\varepsilon}{\theta - 1}$. Furthermore, we have that $s(0)(y) < \frac{1}{n}$ (as otherwise we would get $\sum_{1 \leq y \leq n} s(0)(y) = 1$ so that $s(0)(0) > 0$ could not hold), implying $\frac{\varepsilon}{\theta-1} < \frac{1}{n}$, or $\theta > \varepsilon \cdot n + 1$. This inequality will hold for any $\theta > 1$, if $\varepsilon > 0$ is chosen small enough. ▲

Several aspects of Example 3 are noteworthy. First, the legal limit for how low θ can be with non-selfish play is approached, in the sense that if $\theta < 1$ then selfish play (i.e. $s(x) = n$ for all x) is the unique SE (as is easily inferred on inspection of condition (i) of the Definition). Second, the significant degree of under-reporting present in these SEs is essential to that conclusion, since otherwise $\varepsilon > 0$ could not be selected arbitrarily small, as required in the construction. Third, conceptually, under-reporting may seem esoteric or even anomalous, on intuitive grounds. It is thus natural to wonder whether there are SEs without under-reporting, with bounds on θ that are independent of n. Our next example shows that this is possible although θ cannot be quite as low as with under-reporting:

Example 4: ["sailing-to-the-ceiling"] Consider $s \in S$ such that:

if $x > y$ then $s(x)(y) = 0$,

if $x < n$ then $s(x)(n) = 1 - \varepsilon_n$ where $\varepsilon_n \in (0,1)$,

if $x < y < n$ then $s(x)(y) = (1 - \varepsilon_y) \cdot \prod_{y+1 \leq k \leq n} \varepsilon_k$ where $\varepsilon_y \in (0,1)$.

That is, DM never makes a downward lie, and the probability of cheating via report y is the same for all $x < y$. The SE can be constructed recursively. By the 0-utility-property and using (2), we get

$$T \cdot n - \theta \cdot \sum_{x'} (p(x'|n) \cdot T \cdot [n - x']^{+}) = 0. \tag{5}$$

Since downward lies are ruled out we have $s(n)(n) = 1$, so that $p(x'|n) = \frac{1-\varepsilon_n}{n\cdot(1-\varepsilon_n)+1}$ for all $x' < n$. Plug that into (5), and divide by T, to get

$$n - \theta \cdot \frac{1-\varepsilon_n}{n\cdot(1-\varepsilon_n)+1} \cdot \sum_{x'} [n-x']^+ =$$
$$n - \theta \cdot \frac{1-\varepsilon_n}{n\cdot(1-\varepsilon_n)+1} \cdot \frac{n\cdot(n+1)}{2} = 0. \tag{6}$$

The lhs equals $n - \frac{\theta\cdot n}{2}$ if $\varepsilon_n = 0$, is increasing in ε_n, and is positive for ε_n close enough to 1. Since $\varepsilon_n > 0$, the equation thus has a solution only if $n - \frac{\theta\cdot n}{2} < 0$, or $\theta > 2$. Now, proceeding recursively with $0 < y < n$. By the 0-utility-property and using (2), we get

$$T\cdot y - \theta \cdot \sum_{x'} (p(x'|y)\cdot T\cdot [y-x']^+) = 0. \tag{7}$$

Since downward lies are ruled out we have $s(y)(y) = \prod_{y+1\le k\le n} \varepsilon_k$, so that $p(x'|y) = \frac{(1-\varepsilon_y)\cdot\prod_{y+1\le k\le n}\varepsilon_k}{y\cdot[(1-\varepsilon_y)\cdot\prod_{y+1\le k\le n}\varepsilon_k]+\prod_{y+1\le k\le n}\varepsilon_k} = \frac{1-\varepsilon_y}{y\cdot(1-\varepsilon_y)+1}$ for all $x' < y$. Plug that into (7), and divide by T, to get

$$y - \theta \cdot \frac{1-\varepsilon_y}{y\cdot(1-\varepsilon_y)+1} \cdot \sum_{x'} [y-x']^+ =$$
$$y - \theta \cdot \frac{1-\varepsilon_y}{y\cdot(1-\varepsilon_y)+1} \cdot \frac{y\cdot(y+1)}{2} = 0. \tag{8}$$

The lhs equals $y - \frac{\theta\cdot y}{2}$ if $\varepsilon_y = 0$, is increasing in ε_y, and is positive for ε_y close enough to 1. Since $\varepsilon_y > 0$, the equation thus has a solution only if $y - \frac{\theta\cdot y}{2} < 0$, or $\theta > 2$, the same inequality as applied to ε_n before, so $\theta > 2$ is the general condition for this type of SE to exist. Finally, note that we can rearrange (8), and also (6), to get $\varepsilon_y = 1 - \frac{2}{y\cdot(\theta-2)+\theta}$, where $0 < y \le n$, and then calculate s. ▲

So far we concentrated on SE's with full-support-on-y, but others are possible too. At the extreme, we have the cases of complete pooling: $s \in S$ defined by $s(x) = y$ for all x, for some y. But if $y > 0$, then if θ is high

enough such complete pooling can not be an SE. To see reason in analogy with Observation 3, which already covered the case $y = n$. The only full pooling SE which is robust with respect to increases of θ is that where the pool occurs at 0:

Example 5: ["pooling-at-0"] Consider $s \in S$ defined by $s(x) = 0$ for all x. The easiest way to sustain this as a SE is to assume that $p(0|y) = 1$ for all $y > 0$. Since the utility associated with $y = 0$ is 0, the relevant incentive constraint to rule out a deviation to $y > 0$ is $0 \geq Ty - \theta \cdot 1 \cdot T \cdot y$, or $\theta \geq 1$. The special case of $\theta = 1$ identifies, among all our SE's with non-selfish play, the one which is sustained by the lowest possible θ. ▲

We finally collect the insights scattered above to draw the following insights regarding how the set of equilibria varies with θ :

Observation 5: If $0 \leq \theta < 1$ selfish play ($s(x) = n$ for all x) is the unique SE [cf. the paragraph after Example 3]. If $\theta \geq 1$ there are multiple SE [if $\theta = 1$ this follows from Observation 3 and Example 5; if $\theta > 1$ this follows from Examples 3 & 5]. If $\theta > 1$ there is some SE with full-support-on-y [Example 3]. If $\theta > 2$ there is some SE with full-support-on-y and no under-reporting [Example 4]. For much higher values of θ (and the exact cutoffs increase with n) there are full-support SE that involve cheating-only-at-a-low-x's or lowball-cheating-for-all-x's [Examples 1 & 2]. ■

Amongst all SEs, we feel that the sailing-to-the-ceiling variety highlighted by Example 4 appears as the most robust and perhaps plausible prediction. It has full-support-on-y, exists for any n as long as θ is large enough ($\theta > 2$), works without reliance on downward lies, and is unique as defined.[9] More-

[9] To verify uniqueness, note how ε_y is defined in Example 4 and how this implies uniqueness of $s(x)(y)$. The construction involves that the probability of upward lies from x to any $y > x$ is independent of x. While this assumption seems like a natural benchmark, it could be relaxed in which case uniqueness would be lost although the other properties

over, as we shall see in the next section, its prediction can match F&FH's data quite well.

4 Discussion: Empirical Relevance

How do the predictions of section 3 stand up data? We close our paper by a critical discussion of several related issues.

F&FH's Data Recall that F&FH used a die-rolling paradigm in their study: each participant was asked to privately roll a die and to self-report the outcome to an experimenter, and then got paid based on the report. In our theory, this means that $n = 5$. The predictions of our sailing-to-the-ceiling SE (Example 4) can be eerily similar to F&FH's data (as well as many other studies; cf. Abeler et al's section 1), as shown by Figure 2 where the prediction of our model are generated with $\theta = 3$:

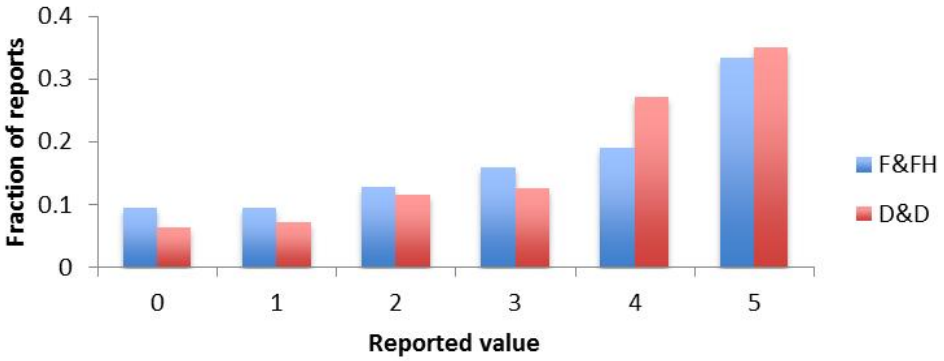


Figure 2, comparing D&D with F&FH.

Heterogeneity: Types vs. Choices F&FH did not observe observe specific die-rolls, but because the knew the underlying die-roll distribution they

listed would stay.

could nevertheless draw inferences regarding the overall nature of cheating across their sample population. Based on their results (as here seen in Figure 2) they offer (p. 536) "three characteristics in the pattern of behavior:

(1) Honest subjects: The fraction of people reporting a payoff of 0 is positive.

(2) Income maximizing subjects: Fraction of people reporting a 5 is above 1/6.

(3) Partial liars: The fraction of people reporting a 4 is above 1/6."

F&FH's phrasing suggests that there exist different types of people – honest, selfish, and partial liars – and of course it may make sense that people are different. Nevertheless, it is intriguing that we can account for the central tendencies in their data without assuming heterogeneity in types. In the SE exhibited in Figure 2, all patterns of choices – (1), (2), and (3) – emerge naturally as part of a particular SE, for a given θ.

Stakes F&FH also examined the effect of stakes on cheating behavior: one treatment tripled all possible payoffs for participants. F&FH found no effect of this change. These findings are consistent with our Observation 1, which says the the set of SE do not depend on our stakes parameter T.[10]

Double-Blind F&FH conduct a double-blind treatment to assess whether partial lies are a result of participants being observed by the experimenter. In their baseline protocol participants roll their die in private, but report the outcome directly to the experimenter who hence can associate each report to an individual. In their double-blind procedure, however, participants are completely anonymous, and their reports cannot be traced back to them. F&FH do not find significant differences in reported outcomes between their normal and double-blind procedures.

On initial inspection this result seems to go against the central premise of our model, that people feel bad if they believe that someone else infers

[10] Mazar et al also found no effect of incentives on cheating behavior, and Abeler et al also report only limited effects (see their Finding 2 on p. 7). However, findings of this sort are not universal, and e.g. Kajackaite & Gneezy report evidence from a design where incentives affected the rate at which people cheat.

that they are cheating. Taking our model at face value we would expect exclusive reporting of the profit-maximizing outcome in F&FH's double-blind procedure since the experimenter no longer serves the role of an audience. However, we offer two important caveats to that conclusion:

First, the effect of anonymity on behavior is a somewhat contested topic. Although F&FH, and also Mazar et al, found no effect of anonymity on cheating, other studies found sizeable effects. For instance, Conrads & Lotz (2015) used a multiple-period design where subjects would flip a coin four times, report the outcomes, and get paid based on the number of reported of "tails." Although partial lying became more prevalent as the channel of communication became less anonymous, dishonesty with respect to extreme and profit-maximizing outcomes became more prevalent with increased anonymity.

Second, we give the following fundamental caveat its own heading:

Being One's Own Audience So far we assumed that our decision maker (DM) cares about the inferences that actual others ("the audience") draw about the extent of his cheating. But DM may alternatively care about the inferences others *would* draw about the extent of his cheating, on observing y, whether or not an actual audience exists. DM might, so to say, internalize a perceived cheating aversion for not behaving such that cheating could be inferred, had he been observed.[11] Our modeling admits such an alternative interpretation, with the added implication that it does not matter whether or not a design is double-blind; our model applies in the same way regardless.

A related intriguing reflection is that if DM is his own audience, in this sense, that may strengthen the presumption that he would play a SE. Our analysis revealed that, as long as $\theta \geq 1$, we had multiple SE. Typically, in game theory, if there are multiple equilibria it is a non-trivial proposition to justify coordination between all players on any particular one. However,

[11] See Akerlof (1983, e.g. p. 57) for some related discussion regarding how, from a parenting point of view, it may be cheaper to inculcate in kids a preference internalizing aspects of honesty that otherwise might plausibly hinge on observability by others.

if DM is his own audience, we have a one-player (psychological) game, so forming equilibrium expectations should be easy.[12]

Self-Concept Maintenance Mazar et al propose that people are often torn between two competing motivations: gaining from cheating and maintaining a self-concept as honest. People solve this dilemma by striking a balance between the financial gain they get from cheating and the maintenance of their positive self-concept, or identity. Mazar et al call this a theory of "self-concept maintenance." F&FH suggest that these ideas are relevant for explaining their findings.

What does it mean to maintain a self-concept as honest? The notion has many facets. We propose that one salient aspect may be to internalize a prefererence for not behaving in such a way that cheating could be inferred, had one been observed. The idea would be that people tie their self-concept perception not to whether or not they actually cheat, but to what impressions they convey, in principle. That concern is then traded off against the material rewards that come with cheating. So, on that interpretation, everything we said under the previous heading ("Being One's Own Audience") might be interpreted as a form of self-concept maintenance, and our model of perceived cheating aversion may be seen as a particular way to formalize Mazar et al's ideas.

Further Tests Some of our SE exist only if n is low. Future experiments could explore how changes in that number affect play (e.g. whether the patterns exhibited by Examples 1 and 2, and Observation 4, occur).

A curious implication is that there are SE where people lie downwards (i.e. report $y < x$). Most existing studies to not speak to the empirical relevance of downward lies, as the experimenter typically observes only subjects' reports, and not the outcomes selected by nature. Future experiments could explore

[12] We get a form of "personal equilibrium" in the sense of Kőszegi (2010).

alternative designs that get around that feature.[13]

Our theory produces testable predictions based on the size of θ. Experiments could tests related propositions, if different subjects could be plausibly deemed to have different values of θ, or if priming may induce changes to θ.

Unlike in standard games, in psychological games information structure across end nodes may matter crucially to predictions (cf. Battigalli & Dufwenberg, pp. 26-27). Our setting can exemplify. We focused on the situation where the audience does not observe x, but one may consider alternatives that alter predictions. For example, if the audience observes both x and y no analog of Observation 2 would hold. In fact, DM's motivation would be as if he had a linear direct cheating cost (cf. section 2 above). Experiments that play around with terminal node information structure may be useful for testing purposes.[14]

References

[1] Abe, N. & J. Greene (2014), "Response to Anticipated Reward in the Nucleus Accumbens Predicts Behavior in an Independent Test of Honesty," *Journal of Neuroscience* 34, 10564-10572

[2] Abeler, J., A. Becker & A. Falk (2014), "Representative Evidence on Lying Costs," *Journal of Public Economics* 113, 96-104.

[3] Abeler, J., D. Nosenzo & C. Raymond (2016), "Preferences for Truth-Telling," mimeo.

[4] Akerlof, G. (1983), "Loyalty Filters," *American Economic Review* 71, 54-63.

[13] In a remarkable study Utikal & Fischbacher (2013) document that *nuns* lie downwards (no one reports the two highest-paying numbers in their die-roll design).

[14] Abeler et al and Gneezy et al run some such experiments and report e.g. that "introducing observability has a strong and significant effect on the distribution of reports" (Abeler et al's Finding 10). Gneezy et al report similar findings (e.g. their Result 7).

[5] Andreoni, J. & D. Bernheim (2009), "Social Image and the 50-50 Norm: A Theoretical and Experimental Analysis of Audience Effects," *Econometrica* 77, 1607-1636.

[6] Arbel, Y., R. Bar-El, E. Siniver & Y. Tobol (2014), "Roll a Die and Tell a Lie – What Affects Honesty?" *Journal of Economic Behavior & Organization* 107, 153-172.

[7] Battigalli, P. & M. Dufwenberg (2007), "Guilt in Games," *American Economic Review*, Papers & Proceedings, 97, 170-176.

[8] Battigalli, P. & M. Dufwenberg (2009), "Dynamic Psychological Games," *Journal of Economic Theory* 144, 1-35.

[9] Bernheim, D. (1994), "A Theory of Conformity," *Journal of Political Economy* 102, 841-877.

[10] Bucciol, A. & M. Piovesan (2011), "Luck or Cheating? A Field Experiment on Honesty with Children," *Journal of Economic Psychology* 32, 73–78.

[11] Cohn, A., E. Fehr & M. Maréchal (2014), "Business Culture and Dishonesty in the Banking Industry," *Nature* 516, 86-89.

[12] Cohn, A., M. Maréchal & T. Noll (2015), "Bad Boys: How Criminal Identity Salience Affects Rule Violation," *Review of Economic Studies* 82, 1289-1308.

[13] Conrads, J. & B. Irlenbusch, R. M. Rilke & G.Walkowitz (2013), "Lying and Team Incentives," *Journal of Economic Psychology* 34, 1-7.

[14] Conrads, J. & S. Lotz (2015), "The Effect of Communication Channels on Dishonest Behavior," *Journal of Behavioral & Experimental Economics* 58, 88-93.

[15] Dai, Z., F. Galeotti & M.-C. Villeval (2016), "Cheating in the Lab Predicts Fraud in the Field. An Experiment in Public Transportations," *Management Science*, forthcoming.

[16] Dieckmann, A., U. Fischbacher, V. Grimm, M. Unfried, V. Utikal & L. Valmasoni (2015), "Trust and Beliefs among Europeans: Cross-Country Evidence on Perceptions and Behavior," Discussion Paper.

[17] Diekmann, A., W. Przepiorka & H. Rauhut (2015), "Lifting the Veil of Ignorance: An Experiment on the Contagiousness of Norm Violations," *Rationality & Society* 27, 309-333.

[18] Dufwenberg, M. & M. Lundholm (2001), "Social Norms and Moral Hazard," *Economic Journal* 111, 506-525.

[19] Ellingsen, T. & M. Johannesson (2008), "Pride and Prejudice: The Human Side of Incentive Theory," *American Economic Review* 98, 990-1008.

[20] Fischbacher, U. & F. Föllmi-Heusi (2013), "Lies in Disguise – An Experimental Study on Cheating," *Journal of the European Economic Association* 11, 525-547.

[21] Fosgaard, T., L. Hansen & M. Piovesan (2013), "Separating Will from Grace: An Experiment on Conformity and Awareness in Cheating," *Journal of Economic Behavior & Organization* 93, 279-284.

[22] Gächter, S. & J. Schulz (2016), "Intrinsic Honesty and the Prevalence of Rule Violations across Societies," *Nature* 531, 496-499.

[23] Geanakoplos, J., D. Pearce & E. Stacchetti (1989), "Psychological Games and Sequential Rationality," *Games & Economic Behavior* 1, 60-80.

[24] Gneezy, U., A. Kajackaite & J. Sobel (2016), "Lying Aversion and the Size of the Lie," unpublished manuscript, UCSD.

[25] Greene, J. & J. Paxton (2009), "Patterns of Neural Activity Associated with Honest and Dishonest Moral Decisions," *Proceedings of the National Academy of Sciences* 106, 12506-12511.

[26] Houser, D., J. A. List, M. Piovesan, A. S. Samek & J. Winter (2016), "On the Origins of Dishonesty: From Parents to Children," *European Economic Review* 82, 242-254.

[27] Houser, D., S. Vetter & Joachim K. Winter (2012), "Fairness and Cheating," *European Economic Review* 56, 1645-1655.

[28] Kajackaite, A. & U. Gneezy (2016), "Lying Costs and Incentives," unpublished manuscript, UCSD.

[29] Kartik, N. (2009), "Strategic Communication with Lying Costs," *Review of Economic Studies* 76, 1359-1395.

[30] Khalmetski, K. & D. Sliwka (2016), "Disguising Lies – Image Concerns and Partial Lying in Cheating Games," unpublished manuscript, University of Cologne.

[31] Kocher, M., S. Schudy & L. Spantig (2016), "I Lie? We Lie! Why? Experimental Evidence on a Dishonesty Shift in Groups," CESifo Working Paper #6008.

[32] Kőszegi, B. (2010), "Utility from Anticipation and Personal Equilibrium," *Economic Theory* 44, 415-444.

[33] Kreps, D. & R. Wilson (1982), "Sequential Equilibrium," *Econometrica* 50, 863-894.

[34] Kroher, M. & T. Wolbring (2015), "Social Control, Social Learning, and Cheating: Evidence from Lab and Online Experiments on Dishonesty," *Social Science Research* 53, 311-324.

[35] Luttmer, E. & M. Singhal (2014), "Tax Morale," *Journal of Economic Perspectives* 28, 149–168.

[36] Mazar, N., O. Amir & D. Ariely (2008), "The Dishonesty of Honest People: A Theory of Self-Concept Maintenance," *Journal of Marketing Research* 45, 633-644.

[37] Muehlheusser, G. A. Roider & N. Wallmeier (2015), "Gender Differences in Honesty: Groups versus Individuals," *Economics Letters* 128, 25-29.

[38] Olken, B. (2015), "Promises and Perils of Pre-Analysis Plans," *Journal of Economic Perspectives* 29, 61–80.

[39] Pascual-Ezama, D., T. Fosgaard, J. Cardenas, P. Kujal, R. Veszteg, B. de Liaño, B. Gunia, D. Weichselbaumer, K. Hilken, A. Antinyan, J. Delnoij, A. Proestakis, M. Tira, Y. Pratomo, T. Jaber-López & P. Brañas-Garza (2014), "Context-Dependent Cheating: Experimental Evidence from 16 Countries," *Journal of Economic Behavior & Organization* 116, 379-386.

[40] Piff, P., D. Stancato, S. Côté, R. Mendoza-Denton & D. Keltner (2012), "Higher Social Class Predicts Increased Unethical Behavior," *Proceedings of the National Academy of Sciences* 109, 4086-4091.

[41] Shalvi, S., J. Dana, M. Handgraaf & C. De Dreu (2011), "Justified Ethicality: Observing Desired Counterfactuals Modifies Ethical Perceptions and Behavior," *Organizational Behavior & Human Decision Processes* 115, 181-190.

[42] Shalvi, S., O. Eldar & Y. Bereby-Meyer (2012), "Honesty Requires Time (and Lack of Justifications)," *Psychological Science* 23, 1264-1270.

[43] Shalvi, S., M. Handgraaf & C. De Dreu (2010), "Ethical Manoeuvring: Why People Avoid Both Major and Minor Lies," *British Journal of Management* 22, S16-S27.

[44] Tadelis, S. (2008), "The Power of Shame and the Rationality of Trust," unpublished manuscript, UC Berkeley.

[45] Utikal, V. & U. Fischbacher (2013), "Disadvantageous Lies in Individual Decisions," *Journal of Economic Behavior & Organization* 85, 108-111.