Available online at www.sciencedirect.com

ScienceDirect

Journal of Economic Theory 175 (2018) 248–264

JOURNAL OF Economic Theory

www.elsevier.com/locate/jet

Lies in disguise – A theoretical analysis of cheating [☆]

Martin Dufwenberg [a,b,c,*], Martin A. Dufwenberg [d]

[a] Department of Economics, University of Arizona, United States
[b] Department of Economics, University of Gothenburg, Sweden
[c] CESifo, Munich, Germany
[d] University of Arizona College of Medicine, United States

Received 8 November 2016; final version received 23 October 2017; accepted 24 January 2018
Available online 31 January 2018

Abstract

We perform a (psychological) game-theoretic analysis of cheating in the setting proposed by Fischbacher and Föllmi-Heusi (2013). The key assumption, referred 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, characterized by three intuitive properties, captures the stylized facts from many experiments (in particular the co-presence of selfish, honest, and partial-lie choices) well.

JEL classification: C72; D82; D83; D91

Keywords: Cheating; Lying; Costs; Audience; Perceived cheating aversion; Psychological game theory

[☆] We started this paper 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, Drew Fudenberg, Navin Kartik, Daniele Nosenzo, Collin Raymond, Joel Sobel, the referees, and several seminar participants for very helpful comments and suggestions.

[*] Corresponding author.

E-mail addresses: martind@eller.arizona.edu (M. Dufwenberg), dufwenberg@email.arizona.edu (M.A. Dufwenberg).

https://doi.org/10.1016/j.jet.2018.01.013

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 rule out, formal sanctions may be helpful. A recent literature, pioneered by Mazar, Amir & Ariely (2008) (MA&A) and Fischbacher & Föllmi-Heusi (F&FH) (2013, but written much earlier), that uses such methods has developed rapidly. 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 across the sample population can be quantified as the experimenters know 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. Others report similar findings. Various explanations have been proposed – F&FH themselves consider lying aversion, caring about lie-credibility, and MA&A's notion of self-concept maintenance.[2]

We propose a new model and explore cheating in settings close to F&FH's (and to signal that link, our paper's title parallels 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 ("DM") issues regarding a random outcome that only DM observes. By issuing a false report DM can mislead with impunity and gain financially. Our central assumption is that DM feels bad to the extent that the audience believes he cheats, 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 undetected, but the audience is smart and draws inferences based on an understanding of DM's 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 (GP&S); Battigalli & Dufwenberg (B&D) 2009). As it turns out, a particular equilibrium, which we show can be characterized by three intuitive properties, allows us to capture the central tendencies of F&FH data remarkably well. There are caveats and twists to that conclusion, including considerations regarding how our theory relates to MA&A's notion of self-concept maintenance. We postpone a discussion until we have introduced our assumptions formally and derived our results.

Section 2 presents the game forms we consider, and some background results. Section 3 introduces our psychological assumption, derives our main prediction, and compares it to F&FH's data. Section 4 contains further analysis. Section 5 discusses testable implications. Section 6

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

[2] For more, see Abeler et al.'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 and Greene (2014), Abeler et al. (2014, 2016), Arbel et al. (2014), Bucciol and Piovesan (2011), Cohn et al. (2014, 2015), Conrads et al. (2013), Conrads and Lotz (2015), Dai et al. (2017), Dieckmann et al. (2015), Diekmann et al. (2015), Fosgaard et al. (2013), Garbarino et al. (2016), Gächter and Schulz (2016), Gneezy et al. (2018), Greene and Paxton (2009), Houser et al. (2012, 2016), Kajackaite and Gneezy (2017), Kocher et al. (2016), Kroher and 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 and Fischbacher (2013).

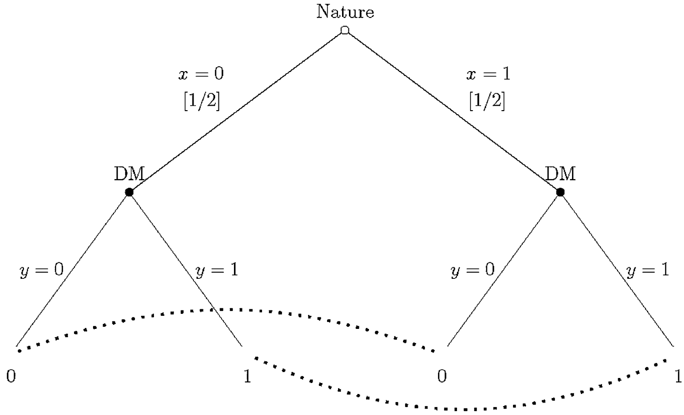

Fig. 1. n = 1 (coin-flip).

compares our model to related theoretical work, in particular by Gneezy, Kajackaite & Sobel (2018) (GK&S), Khalmetski & Sliwka (2017) (K&S), and Abeler, Nosenzo & Raymond (2016) (AN&R) who also explore belief-dependent preferences in F&FH's setting. Section 7 concludes, focusing on interpretations.

2. Preliminaries

Game forms Nature randomly draws $x \in \{0, ..., n\}$, $n \geq 1$, with probability $\pi_x \in (0, 1)$, $\sum_x \pi_x = 1$. 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 $s : \{0, ..., n\} \rightarrow \Delta\{0, ..., n\}$, so $s(x)(y)$ denotes the probability that 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&FH who focus on the die-roll case: x is drawn from a uniform distribution with $n = 5$.[3] Others studied the coin-flip case ($n = 1$)[4]; the associated game-tree is illustrated in Fig. 1.

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 forms 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

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

[4] Bucciol and Piovesan (2011) may have been first to do this.

$y = n$ regardless of x. If $\theta = 1$ he may choose any element of $\Delta\{x, ..., n\}$. If $\theta > 1$ he chooses $y = x$.[5]

Direct cheating costs are comparable to lying costs that depend (e.g. linearly) on the size of the lie (here $|y - x|$); see 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

In this section we define utility & equilibrium (3.1), prove results that imply partial cheating (3.2), derive our main prediction based on three auxiliary assumptions (3.3), and compare it to F&FH's data (3.4).

3.1. Utility & solution concept

We first introduce our central psychological assumption: *DM feels bad to the extent that the audience believes he is cheating*. It may be plausible to combine this sentiment with direct cheating costs, but in order to highlight the insights that come from our main assumption we disregard direct cheating costs except as a yardstick for comparison.

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

$$\begin{aligned} &T \cdot y - \theta \cdot \sum_{x'} (p(x'|y) \cdot T \cdot [y - x']^+) = \\ &= T \cdot y - \theta \cdot \sum_{x' < y} (p(x'|y) \cdot T \cdot [y - x']) \end{aligned} \tag{2}$$

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 GP&S and B&D (2009).[6] The utility is independent of x. DM cares about his image, not about cheating per se. We say that (2) captures *perceived cheating aversion*.

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 two conditions holds:

[5] To escape this all-or-nothing implication, one may consider convex costs that yield partial lies of specific step-length, that 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 $y = \max\{x + 1, n\}$.

[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.

(i) $s(x)(y) > 0 \Rightarrow y \in \arg\max_z (T \cdot z - \theta \cdot \sum_{x'<z}(p(x'|z) \cdot T \cdot [z - x']))$,
(ii) $\sum_x s(x)(y) > 0 \Rightarrow p(x'|y) = \frac{\pi_{x'} \cdot s(x')(y)}{\sum_x \pi_x \cdot s(x)(y)}$.

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 beliefs. If y is a choice DM will not make ($\sum_x s(x)(y) = 0$) then no restrictions apply (except $p(x'|y) \in [0, 1]$, $\sum_{x'} p(x'|y) = 1$). The SE-terminology is justified in that predictions coincide with Kreps and Wilson's (1982) classic notion if $\theta = 0$, and with B&D's (2009) extension of SE to psychological games if $\theta \geq 0$.[7]

3.2. Partial lying

The following three observations, each interesting in its own right, jointly imply an important general insight regarding SE play:

Observation 1. (*a*) At least one SE always exist. (*b*) The set of SE 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 SE-Definition. □

The next two results rule out honest reporting and, for high enough θ, full-blown cheating:

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

Proof. Suppose $s(x) = x$ for all x and that (ii) of the SE-Definition holds. If $x' \neq y$ then $p(x'|y) = \frac{0}{n \cdot 0 + 1} = 0$, so (1) equals 0. 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. Hence s is not an SE. □

Intuitively, if honesty were an equilibrium the audience would believe that $x = y$, so cheating at $y = n > x$ would go undetected. However, as we show next, for high enough θ, pooling on n cannot be a SE either; strong perceived cheating would result and DM escapes that impression by deviating to $y = 0$:

Observation 3. There exist $\widehat{\theta}((\pi_x)_{x \leq n}) > 1$ such that $s \in S$ defined by $s(x) = n$ for all x is a SE iff $\theta \leq \widehat{\theta}((\pi_x)_{x \leq n})$. If x is drawn from a uniform distribution ($\pi_x = \frac{1}{n+1}$ for all x) then $\widehat{\theta}((\pi_x)_{x \leq n}) = 2$.

Proof. If s is a SE, DM's utility cannot be negative as he could always get 0 by choosing $y = 0$ (no cheating is possible then). Moreover, the audience has correct initial beliefs, so by (ii) of the SE-Definition we get $p(x'|n) = \pi_{x'}$ for all $x' \in X$. Using (2), we get

[7] B&D (2009) 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 $p(x'|y)$.

$$T \cdot n - \theta \cdot \sum_{x'<n} (\pi_{x'} \cdot T \cdot [n - x']) \geq 0, \tag{3}$$

and, rearranging, $\theta \leq n / \sum_{x'<n} (\pi_{x'} \cdot [n - x'])$. Define and note: $\widehat{\theta}((\pi_x)_{x \leq n}) = n / \sum_{x'<n} (\pi_{x'} \cdot [n - x']) > 1$.

To see that $\theta \leq \widehat{\theta}((\pi_x)_{x \leq n})$ is sufficient for s to be a SE, set $p(0|y) = 1$ for all $y < n$. (3) shows that the utility of choosing $y = n$ is non-negative. By comparison, using (2), we see that the utility of choice y equals $T \cdot y - \theta \cdot 1 \cdot T \cdot [y - 0] = T \cdot y \cdot (1 - \theta)$ which is non-positive if $\theta \in [1, \widehat{\theta}((\pi_x)_{x \leq n})]$; hence there is no profitable deviation from n to $y < n$ in this case. Finally suppose that $\theta < 1$. Since the lhs of (3) is larger than $T \cdot n \cdot (1 - \theta) > 0$, and $T \cdot y \cdot (1 - \theta)$ is increasing in y, we can again draw the conclusion that no profitable deviation to $y < n$ is available.

If x is drawn from a uniform distribution then (3) becomes $Tn - \theta \cdot T \cdot \frac{n}{2} \geq 0$, implying $\theta \leq 2 = \widehat{\theta}((\pi_x)_{x \leq n})$. □

Observations 1(a), 2, and 3 combine to imply that for high enough θ SE exist but must exhibit partial lying, in stark contrast with the all-or-nothing conclusions associated with linear direct costs of cheating of section 2. This insight paves the way for the analysis to follow.

3.3. Main prediction [“sailing-to-the-ceiling”]

The partial lying prediction of section 3.2 comes in many guises, as there are multiple, qualitatively different SE (see section 4.1). However, three intuitive properties imply that a unique SE is viable:

- s has *full-support-on-y* if $\sum_x s(x)(y) > 0$ for all y.
- s exhibits *no-downward-lies* if $s(x)(y) = 0$ for all $y < x$.
- s exhibits *uniform-cheating* if $s(x)(y)$ is the same for all $x < y$.

The first property is empirically well grounded: in experiments all y-choices occur. The other two properties may be somehow focal when subjects reason: downward lying might feels counterintuitive while uniform-cheating could come natural as the utility of $y > x$ doesn’t depend on x (see section 3.1). In any case, as we show in section 3.4, the implied SE can exhibit close resemblance with data.

Proposition. *If s is a SE with full-support-on-y that exhibits no-downward-lies and uniform-cheating, then s is the only SE with those properties. There exists $\widehat{\theta}((\pi_x)_{x \leq n}) > 1$ such that said SE exists iff $\theta > \widehat{\theta}((\pi_x)_{x \leq n})$. If x is drawn from a uniform distribution ($\pi_x = \frac{1}{n+1}$ for all x) then $\widehat{\theta}((\pi_x)_{x \leq n}) = 2$.*

Our proof is constructive. It makes critical use of the insight that if s is a SE with full-support-on-y then the utility following any choice must be the same. Otherwise, since DM’s utility does not depend on x, he would have a profitable deviation to whatever y gave higher utility. In fact, the utility must equal 0, since this is what that DM gets if he chooses $y = 0$ (no cheating is possible then). Say that s then has the 0-*utility-property.*

Proof. Let $s \in S$ be a SE 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 $\varepsilon_y \in (0, 1)$.

I.e., s exhibits no-downward-lies, uniform-cheating, and (since $\varepsilon_y \in (0, 1)$ and $\varepsilon_1 > 0$ implies $s(0)(0) > 0$) full-support-on-y.

By the 0-utility-property, using (2), we get

$$T \cdot n - \theta \cdot \sum_{x' < n} (p(x'|n) \cdot T \cdot [n - x']) = 0. \tag{4}$$

Since $s(n)(n) = 1$ and $s(x)(n) = 1 - \varepsilon_n$ for all $x < n$, using (ii) of the SE-Definition, we get

$$\begin{aligned} p(x'|n) &= \frac{\pi_{x'} \cdot (1 - \varepsilon_n)}{\sum_{x<n}(\pi_x \cdot (1 - \varepsilon_n)) + \pi_n \cdot 1} = \\ &= \frac{\pi_{x'}}{(1 - \pi_n) + \frac{\pi_n}{(1-\varepsilon_n)}} \end{aligned} \tag{5}$$

for all $x' < n$. Plug (5) into (4), and divide by T, to get

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

The lhs of (6):
equals $n - \theta \cdot \sum_{x'<n}(\pi_{x'} \cdot [n - x'])$ if $\varepsilon_n = 0$,
is increasing in ε_n,
and is positive for ε_n close enough to 1.

Since $\varepsilon_n > 0$, (6) thus has a solution iff $n - \theta \cdot \sum_{x'<n}(\pi_{x'} \cdot [n - x']) < 0$, or $\theta > n / (\sum_{x'<n} \pi_{x'} \cdot [n - x']) > 1$. Note that the solution is unique.

Proceed recursively with $0 < y < n$. By the 0-utility-property, using (2):

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

No-downward-lie implies $s(y)(y) = \prod_{y+1 \leq k \leq n} \varepsilon_k$. Since $s(x)(y) = 1 - \varepsilon_y$ for all $x < y$, we get

$$\begin{aligned} p(x'|y) &= \frac{\pi_{x'} \cdot (1 - \varepsilon_y) \cdot \prod_{y+1 \leq k \leq n} \varepsilon_k}{\sum_{x<y}(\pi_x \cdot (1 - \varepsilon_y) \cdot \prod_{y+1 \leq k \leq n})\varepsilon_k + \pi_y \cdot \prod_{y+1 \leq k \leq n} \varepsilon_k} = \\ &= \frac{\pi_{x'}}{(\sum_{x<y} \pi_x) + \frac{\pi_y}{(1-\varepsilon_y)}} \end{aligned} \tag{8}$$

for all $x' < y$. Plug (8) into (7), and divide by T, to get

$$y - \theta \cdot \sum_{x' < y} (\frac{\pi_{x'}}{(\sum_{x<y} \pi_x) + \frac{\pi_y}{(1-\varepsilon_y)}} \cdot [y - x']) = 0. \tag{9}$$

The lhs of (9):
equals $y - \theta \cdot \sum_{x'<y}(\frac{\pi_{x'}}{(\sum_{x<y} \pi_x) + \pi_y} \cdot [y - x'])$ if $\varepsilon_y = 0$,
is increasing in ε_y,
and is positive for ε_y close enough to 1.

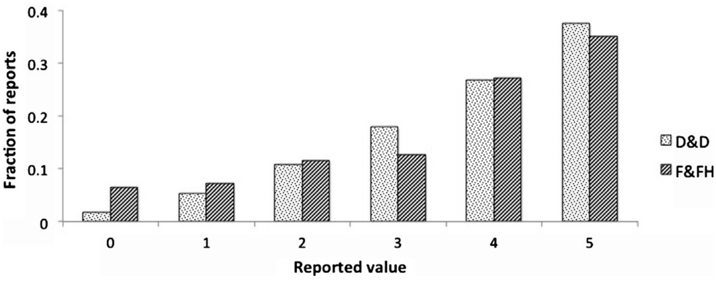

Fig. 2. D&D theory vs F&FH data.

Since $\varepsilon_y > 0$, (9) has a solution iff $y - \theta \cdot \sum_{x'<y}(\frac{\pi_{x'}}{(\sum_{x<y}\pi_x)+\pi_y} \cdot [y - x']) < 0$, or $\theta > y/\sum_{x'<y}(\frac{\pi_{x'}}{\sum_{x \le y}\pi_x} \cdot [y - x']) > 1$. The solution is unique.

Inspecting the conditions on θ needed for solutions of (6) & (9) to exist (mentioned after (6) & (9)), one infers that an SE exists iff:

$$\theta > \widehat{\theta}((\pi_x)_{x \le n}) = \max_{y \in \{1,...,n\}} y/\sum_{x'<y}(\frac{\pi_{x'}}{\sum_{x \le y}\pi_x} \cdot [y - x']). \tag{10}$$

If (10) holds, since (6) and (9) uniquely define $\varepsilon_y \in (0, 1)$ for $y \in \{1, ..., n\}$, the SE s is uniquely defined while satisfying all the desired properties.

If x is drawn from the uniform distribution we get $\theta > \widehat{\theta}((\pi_x)_{x \le n}) = 2$. To see this, plug $\pi_x = \pi_{x'} = \frac{1}{n+1}$ for all x, x' into the rhs of (10), which then equals $y/(y \cdot \frac{\frac{1}{n+1}}{(y+1)\cdot\frac{1}{n+1}} \cdot \frac{y+1}{2}) = 2$. □

It is useful to have a name for the SE highlighted in the Proposition. It involves that if DM observes x then he reports each $y > x$ with positive probability; we suggest *sailing-to-the-ceiling* as an apt monicker.

The critical value $\widehat{\theta}((\pi_x)_{x \le n})$, defined in (10), depends on the distribution $(\pi_x)_{x \le n}$, and may be affected by n in interesting ways (although n is irrelevant if the distribution is uniform). We return to this topic in section 4.2.

3.4. The proposition vs. F&FH's data

How does the sailing-to-the-ceiling SE of the Proposition stand up to data? Most studies assume that x is drawn from a uniform distribution. This simplifies computation. Namely, plug $\pi_x = \frac{1}{n+1}$ for all x into (6) & (9). For $0 < y \le n$ we get

$$\begin{aligned} & y - \theta \cdot \frac{1-\varepsilon_y}{y \cdot (1-\varepsilon_y) + 1} \cdot \frac{y \cdot (y+1)}{2} = 0 \\ \Longleftrightarrow \quad & (1 - \varepsilon_y) = \frac{2}{y \cdot (\theta - 2) + \theta}. \end{aligned} \tag{11}$$

Focus on F&FH's die-roll setting: $n = 5$. Using (11), we can compute the SE, which can be eerily similar to F&FH's data (and other studies; see AN&R). This is illustrated in Fig. 2 where the prediction is generated with $\theta = 3$.[8]

See section 5 for more discussion of experimental tests.

[8] Using (11), we get $(1 - \varepsilon_5) = \frac{2}{5(3-2)+3} = \frac{1}{4}$, so $\sum_{x \le 5} \pi_x \cdot s(x)(5) = 5(\frac{1}{6} \cdot \frac{1}{4}) + \frac{1}{6} \cdot 1 = \frac{3}{8}$, etc.

4. Further analysis

4.1. Other SE

The uniqueness of the sailing-to-the-ceiling SE of section 3.3 relies on s exhibiting no-downward-lies, uniform-cheating, and full-support-on-y. Relaxing these requirements, mostly one at a time, we now illustrate how other SE come alive. Unless we say otherwise, we assume that x is drawn from a uniform distribution.

The proof of the Proposition defined a unique $\varepsilon_y \in (0,1)$ for all $y \in \{1, ..., n\}$. Uniqueness of SE is lost if we give up uniform-cheating and replace $\varepsilon_y \in (0,1)$ by $\varepsilon_{yx} \in (0,1]$ for all $x < n$, $y \geq 1$. Following the spirit of the Proposition, *mutatis mutandis*, a host of SE can be constructed. However, n limits the SE-possibilities for some classes of strategies, as the following two examples + observation show. The SE described exhibit full-support-on-y and no-downward-lies, but not uniform-cheating.

Example 1. [“rotten-0”] Consider $s \in S$ such that:
$s(0)(y) > 0$ for all y,
$s(x) = x$ for all $x \neq 0$,.

Cheating occurs only when $x = 0$, and involve all $y > 0$. For s to be a SE, by the 0-utility-property, 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 (with π_x uniform) $p(0|y) = \frac{s(0)(y)}{s(0)(y)+1}$ we get $\frac{1}{\theta} = \frac{s(0)(y)}{s(0)(y)+1}$ or $s(0)(y) = \frac{1}{\theta-1}$. Furthermore, $s(0)(y) < \frac{1}{n}$ (as otherwise $s(0)(0) > 0$ 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 $y \in \{x, x+1\}$,
$s(n) = n$.

Cheating occurs with positive probability for all $x < n$, but involves only the smallest exaggeration of one-unit. The SE can be constructed recursively. By the 0-utility-property, and using (2), for $y > 0$ we get $T \cdot y - \theta \cdot p(y-1|y) \cdot T \cdot 1 = 0$, or $p(y-1|y) = \frac{y}{\theta}$. By (ii) of the SE-Definition (and π_x uniform) we have $p(y-1|y) = \frac{s(y-1)(y)}{s(y-1)(y)+s(y)(y)}$. Hence

$$\frac{y}{\theta} = \frac{s(y-1)(y)}{s(y-1)(y) + s(y)(y)}, \tag{12}$$

or $s(y-1)(y) = \frac{y \cdot s(y)(y)}{\theta - y}$. 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 (12), that $s(y-1)(y)$ is well-defined for all $y \in \{1, ..., n\}$, and that $0 < s(y-1)(y) < s(n-1)(n)$.[9]

The SE in Examples 1 & 2 display striking idiosyncratic cheating patterns. However, for a given θ, a large enough n undermines their existence. The respective necessary conditions, $\theta > n + 1$ and $\theta > 2n$, cannot hold if n is large enough. To get a better feel for why, note that in

[9] 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 (12) is increasing in y, and that in the denominator of the rhs of (12) it holds that $s(y)(y) < s(n)(n) = 1$.

Example 1, to satisfy the 0-utility-property, all cheating is done when $x = 0$. With $s(x)(x) = 1$ for all $x > 0$, to achieve $p(0|x) = \frac{1}{\theta}$, $s(0)(x)$ must reach a critical strictly positive level for all $x > 0$, independently of n. This becomes impossible for high n as there is only a unit of probability to play with. In Example 2, to achieve the 0-utility for $y = n$, all cheating is done for $x = n - 1$. The amount of cheating when $x = n - 1$ and $y = n$ is small ($y - x = n - (n - 1) = 1$), so 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 $l \in \mathbb{N}_0$, and let x be drawn from a uniform distribution. (a) Consider $s \in S$ with full-support-on-y such that $s(x) = x$ for all $x > l$. If n is large enough then s is not a SE. (b) Consider $s \in S$ with full-support-on-y such that $s(n) = n$ while for all $x < n$ we have $s(x)(y) > 0$ only if $y \in \{x, x + l\}$. If n is large enough then s is not a SE.

We omit the proof. The essence mirrors the intuition highlighted via Examples 1 & 2: For a given DM, for whom θ reflects his given personality, neither cheating-only-at-a-low-x's nor lowball-cheating-for-all-x's can be sustained as SE, for large enough n.[10]

We next construct a SE with full-support-on-y and uniform-cheating, which however involves downward-lies.

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 $y > 0$. 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 (with π_x uniform) it holds that $p(0|y) = \frac{s(0)(y)}{s(0)(y)+\varepsilon}$ 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 condition $\theta > 1$ is less stringent than the corresponding condition $\theta > \widehat{\theta}((\pi_x)_{x\leq n}) = 2$ for the sailing-to-the-ceiling SE (with a uniform distribution). Second, the legal limit for how low θ can be with non-selfish play is approached, as if $\theta < 1$ 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 SE-Definition). Third, the under-reporting present in a crowded-floor SE is essential to that conclusion; otherwise $\varepsilon > 0$ could not be selected arbitrarily small, as required in the construction.

So far all exhibited SE involve behavior that depends on x. Is it possible to have a SE such that $s(x)(y)$ is independent of x? The answer is largely negative. To see this, consider a SE with full-support-on-y, with x drawn from a uniform distribution. Since $s(x)(y) > 0$ is independent

[10] Of course, one may also prove the flip-side result that for a given n, if θ is high enough s as described in Observation 4 may be a SE. This could be interesting to know if one fixes a game and asks for whom (among folks with different θ) s as defined could be a SE.

of x, if s were an SE we'd get $p(x'|y) = \frac{1}{n+1}$ for all x', y, and by the 0-utility property, using (2), $T \cdot y - \theta \cdot y \frac{1}{n+1} \cdot T \cdot \frac{y+1}{2} = 0$, which cannot hold for different values of y.

If we do not impose full-support-on-y, there are SE with complete pooling (which of course satisfy that $s(x)(y)$ is independent of x). If $s(x) = n$ for all x we get uniform-cheating and no-downward-lies, and a SE iff $\theta \in [1, 2]$ (as follows from Observation 3 + the remark after Example 3). If $0 < y < n$ and $s(x) = y$ for all x, there is downward lying, and s cannot be an SE if θ is high enough. To see that, reason in analogy with Observation 3, which already covered the case $y = n$. The only SE with full pooling which is robust with respect to increases of θ is that where the pool occurs at 0. It is valid for all distributions $(\pi_x)_{x \leq n}$.

Example 4. ["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.

4.2. $\widehat{\theta}((\pi_x)_{x \leq n})$ and n

As seen in (10), the number $\widehat{\theta}((\pi_x)_{x \leq n})$, critical to the existence of the sailing-to-the-ceiling SE, depends on $(\pi_x)_{x \leq n}$. Suppose we consider some natural class of such distributions, parameterized by n. One may wonder: if one fixes θ, does the sailing-to-the-ceiling SE cease to exist if n is high enough (per analogy with Observation 4)? Or, more generally, how does $\widehat{\theta}((\pi_x)_{x \leq n})$ vary with n? The possibilities abound:

Observation 5. There are classes of distributions of x, parameterized by n, such that if n increases then $\widehat{\theta}((\pi_x)_{x \leq n})$ (defined by (10)) (i) decreases, (ii) is constant, (iii) increases within bound, or (iv) increases without bound.

Proof. (i) Suppose that $\pi_n = \frac{k}{n+1}$ for some $k \in (1, 2)$ while $\pi_x = \frac{1-\pi_n}{n}$ for all $x < n$. Recall the SE-construction in the proof of the Proposition, and consider the following key component of (10):

$$y / \sum_{x' < y} \left(\frac{\pi_{x'}}{\sum_{x \leq y} \pi_x} \cdot [y - x'] \right). \tag{13}$$

If $y < n$ then (13) equals 2. To see this, plug $\pi_x = \pi_{x'} = \frac{1-\pi_n}{n}$ for all x, x' into (13), which then equals $y/(y \cdot \frac{\frac{1-\pi_n}{n}}{(y+1) \cdot \frac{1-\pi_n}{n}} \cdot \frac{y+1}{2}) = 2$. If we consider (13) for $y = n$, noting that $\sum_{x \leq n} \pi_x = 1$, we get $n/((1 - \pi_n) \cdot \frac{n+1}{2}) = \frac{n}{n+1-k} \cdot 2 > 2$. Hence (by (10)), $\widehat{\theta}((\pi_x)_{x \leq n}) = \frac{n}{n+1-k} \cdot 2$, which is decreasing in n.

(ii) Suppose that $\pi_n = \frac{k}{n+1}$ for some $k \in (0, 1]$ while $\pi_x = \frac{1-\pi_n}{n}$ for all $x < n$. If $y < n$ then (13) equals 2 (as in (i)). If we consider (13) for $y = n$, we get $\frac{n}{n+1-k} \cdot 2$, just as in (i) except that now $\frac{n}{n+1-k} \cdot 2 \leq 2$. Hence $\widehat{\theta}((\pi_x)_{x \leq n}) = 2$ for all n. (This generalizes the result we had for the uniform distribution in the Proposition.)

(iii) Suppose that $\pi_n = \pi \in (0, 1)$ while $\pi_x = \frac{1-\pi}{n}$ for all $x < n$. If n is small enough that $\pi < \frac{1}{n+1}$, case (ii) above applies so we get $\widehat{\theta}((\pi_x)_{x \leq n}) = 2$. If n is large enough that $\pi > \frac{1}{n+1}$ then for $y < n$ (13) equals 2 (as in (i) and (ii)) and if we consider (13) for $y = n$, we get $n/((1 -$

$\pi) \cdot \frac{n+1}{2}) = \frac{n}{n+1} \cdot \frac{2}{1-\pi} > 2$. Hence $\widehat{\theta}((\pi_x)_{x\leq n}) = \frac{n}{n+1} \cdot \frac{2}{1-\pi}$, which is increasing in n but bounded from above by $\frac{2}{1-\pi}$.

(iv) Suppose that $\delta \in (0, \frac{1}{2})$ and that $\pi_x = \delta^{n-x}$ for $x < n$, while $\pi_n = 1 - \sum_{x<n} \pi_x$. Note that $\pi_n > 0$ for all n. $\widehat{\theta}((\pi_x)_{x\leq n})$ can be no smaller than (13) with $y = n$, so $\widehat{\theta}((\pi_x)_{x\leq n}) \geq n / \sum_{x'<n}(\delta^{n-x'} \cdot [n - x'])$. Furthermore, the following is true:

$$\frac{n}{\sum\limits_{x'<n} (\delta^{n-x'} \cdot [n-x'])} > \frac{n}{\sum\limits_{x'<n} (\delta^{n-x'} \cdot 2^{n-x'})} > \frac{n}{\sum\limits_{k\geq 0} (2\delta)^k} = n \cdot (1 - 2\delta). \tag{14}$$

The first inequality holds since $2^{n-x'} > n - x'$ for $0 \leq x' < n$, the second by adding positive terms in a denominator. The equality holds since $0 < 2\delta < 1$. Combining inequalities, we get $\widehat{\theta}((\pi_x)_{x\leq n}) > n \cdot (1 - 2\delta)$. Noting that $\lim_{n\to\infty} n \cdot (1 - 2\delta) = \infty$, the conclusion follows. □

5. Testable implications

We already illustrated (in section 3.4) that the sailing-to-the-sailing SE, highlighted by the Proposition, can match data in F&FH's setting well. We now discuss other experimental tests that could be done.

Our model's parameters – n, T, θ, $(\pi_x)_{x\leq n}$ – relate in testable ways to the sailing-to-the ceiling prediction. In section 3.4 we looked at the case where $n = 5$ and $\theta = 3$. Further tests could target out-of-sample predictions for other values of n.[11] T is predicted not to matter at all.[12] As regards θ, experiments could test related implications, if different subjects could plausibly be deemed to have different values, or if priming might induce changes. As regards $(\pi_x)_{x\leq n}$, we concentrated on the case where x is drawn uniformly, but the SE of section 3.3 was defined more generally. Consider, for example, the cases, considered by GK&S, where $0 < \pi_n \leq \pi_x = \pi$ for all $x < n$. Inspecting (6), one sees that if π_n is reduced then $s(x)(n)$ is reduced as well, in line with GK&S's data.

Tests related to n may have a special interest when considering non-uniform distributions, related to the results of section 4.2. For example, if we consider the distribution given in Observation 5(iv) ($\delta < \frac{1}{2}$, $0 < \pi_x = \delta^{n-x}$ for $x < n$, $\pi_n = 1 - \sum_{x<n} \pi_x$), then if n is large enough the sailing-to-the-ceiling SE is no longer viable (for a given θ).

While we presented the sailing-to-the ceiling SE as our main prediction, experimental tests may relate also to other SE. For example, some SE exist only if n is low; one could explore whether the patterns exhibited by Examples 1 & 2, and Observation 4, occur. Another curious aspect of some SE (e.g. Example 3) is that people lie downwards (i.e. report $y < x$). Most existing studies do not speak to the empirical relevance, as the experimenter typically observes only DM's choice y, and not Nature's choice x. Perhaps one could explore alternative designs that somehow get around that feature.[13]

[11] Using (11), we get $(1 - \varepsilon_y) = \frac{2}{y(\theta-2)+\theta} = \frac{2}{y+3}$ if $\theta = 3$, so $\sum_x \pi_x \cdot s(x)(n) = n(\frac{1}{n+1} \cdot \frac{2}{n+3}) + \frac{1}{n+1} \cdot 1 = \frac{3}{n+3}$, e.g. $\frac{1}{34}$ if $n = 99$.

[12] F&FH examined the effect of stakes: one treatment tripled all possible payoffs. They found no effect of this change, consistent with our Observation 1(b). MA&A and AN&R similarly report only limited effects. 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.

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

Our analysis does not depend on the audience's payoff, which is why we didn't specify it. Experiments could test whether, indeed, cheating behavior is independent of how reports affect the audience (or others).[14]

Finally, we note that in psychological games information structure across end nodes may matter to predictions (see B&D 2009, pp. 26–27). Our setting can exemplify. We focused on the situation where the audience does not observe x, but one may consider alternatives. 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 (see section 2). Experiments that play around with terminal node information structure may be useful for testing purposes.[15]

6. Related work

F&FH's setting GK&S and K&S consider a form of belief-dependent utility, namely that DM experiences disutility proportional to the probability that the audience assigns that he lied. They additionally let utilities reflect direct lying costs, but let us temporarily disregard those and compare (2) to the following utility reflecting dislike of being deemed a liar:

$$T \cdot y - \theta \cdot \sum_{x' \neq y} p(x'|y). \tag{15}$$

While (2) and (15) capture similar psychological intuitions, the absence of $T \cdot [y - x']^+$ in (15), relative to (2), implies the following behavioral difference: Recall how (in section 3.3) we constructed an SE with full-support-on-y, viable for $\theta \geq 2$ and any n, if x is drawn from a uniform distribution. (15) does not allow an analogous SE. To see this, note first that $0 \leq \sum_{x' \neq y} p(x'|y) \leq 1$. The utility (15) of choosing $y = 0$ is thus bounded from above by number $T \cdot 0 - \theta \cdot 0 = 0$, and the utility of choosing $y = n$ is bounded from below by number $T \cdot n - \theta$. If T or n is large enough, $T \cdot n - \theta > 0$, an inequality that rules out an SE with full-support-on-y since all choices of y (including $y = 0$ and $y = n$) must give the same utility in said SE.[16]

The models of GK&S and K&S augment (15) to incorporate also a fixed direct cost of lying, incurred iff $y \neq x$, drawn from a distribution with support on such high values that (regardless of n) some types never lie. This guarantees that any SE must have full-support-on-y, and that $p(x|x) > 0$. Furthermore, they prove a striking uniqueness result. Namely, a SE with the following properties (which notably are not shared by our sailing-to-the-ceiling SE) will be played: There is a k such that if $x < k$ DM lies with positive probability and then reports some $y > k$. Moreover, if $x > k$ DM does not lie. The prediction is unique as regards the implied distribution over y. For various parameter-constellations, such SE generate predictions that resemble various data sets, including F&FH's.[17]

[14] In most experiments the audience is the experimenter, who keeps what DM does not claim. In some studies, including a treatment by F&FH, another (dummy player) subject is given what DM does not claim. F&FH found no significant effect in that variation, although some subsequent studies did (see AN&R).

[15] AN&R and GK&S consider some related treatments, and report e.g. that "introducing observability has a strong and significant effect on the distribution of reports" (AN&R, Finding 10). GK&S report similar findings (e.g. Result 4).

[16] To gain further intuition, reflect on the role that factor $T \cdot [y - x']^+$ plays in (2), as compared to (15) where it is absent. It allows, in (2), that if perceived cheating is high, the overall utility is reduced in proportion. (15) has no analogous feature; perceived lying is at most "probability 1," which isn't enough to outweigh the material rewards of lying, if T or n are large enough.

[17] K&S set $T = 1$. GK&S's model is more general, with report-values v_y increasing in y, and lying costs that depend in more detail on x and y; see their paper for details.

AN&R discuss how a variety of theories relate to F&FH's setting, and initiate discussion of how incomplete information about parameters like θ shape predictions.[18] We refer to their text for more on that topic, as well as coverage of models that involve no component of belief-dependent utility.

Psychological game theory Beyond F&FH's setting, our model is related to work which incorporates belief-dependent utility. Particularly relevant are applications where a player is somehow concerned with another's belief-dependent "opinion" at end nodes. For example, Dufwenberg and Lundholm (2001) study individuals with privately observed "talent," who choose "effort," observed by "neighbors," who bestow "social respect" (based on a "social norm" that more talented individual should exert more effort). Their setup is structurally parallel to ours: Nature makes a choice (talent vs. x) which a player (individual vs. DM) observes before making a choice (effort vs. y) which is observed by another (neighbor vs. audience) who forms beliefs (social respect vs. perceived cheating) that the first player cares about. Utilities differ, but the game-theoretic structure is comparable.

Similar parallels can be drawn to models where a player's preference parameter is determined by Nature, and the player cares about another's beliefs regarding that parameter, which the player can influence by some choice. See Bernheim (1994) on "status," Ellingsen and Johannesson (2008) on "pride & prejudice," Andreoni and Bernheim (2009) on "social image," and Blume, Lai and Lim (2016) on "randomized response." Tadelis (2008) study of "shame" is also related. The issue is not that player i cares about j's belief about a choice by Nature, but rather that i cares about j's beliefs about i's choice. However, j's belief is affected by the realization of a chance move unobservable to j, in analogy with how in our study DM cares about the audience's ex post beliefs (which are affected by x which is hidden from the audience's view).

Similar information issues arise in B&D's (2007) model of guilt-from-blame, where however the belief-dependent part of i's utility does not concern j's beliefs about i's choice but rather j's beliefs regarding the extent to which i believes his choices give j less than j expects. That idea can be applied to any game form. One may wonder if perceived cheating aversion can be similarly extended. However, that sentiment concerns reporting of private information and associated opportunities for cheating. It can only be meaningfully applied to game forms that allow such an interpretation.

The economic issues addressed are quite different across these studies. However, since they deal with belief-dependent utility, all are exercises that fit within the framework of psychological game theory.

7. Discussion

We wrap up with a few remarks that mainly interpret our model and results.

Heterogeneity: types vs. choices F&FH did not observe specific die-rolls, but because they knew the underlying die-roll distribution they could draw inferences regarding the overall nature of cheating across their sample population. Based on their results (as seen in Fig. 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.

[18] In section 7 below, under the heading "Being one's own audience," we argue that assuming complete information about θ may not be restrictive.

(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 many types of people – honest, selfish, and partial liars – and of course it makes sense that people are different. Yet, it is intriguing that we can account for the central tendencies in their data without assuming heterogeneity in types. In the SE exhibited in Fig. 2, all three patterns of choices emerge naturally for a given θ.

Double-blind F&FH conduct a double-blind treatment to assess whether partial lies are a result of participants being observed. 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, not even their reports can 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 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 caveats to that conclusion:

First, the effect of anonymity on behavior is a somewhat contested topic. Although F&FH, and also MA&A, found no effect of anonymity on cheating, other studies found effects. For instance, Conrads and 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 "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 on observing y, whether or not an actual audience exists. DM might, so to say, internalize a preference for not behaving such that cheating could be inferred, had he been observed.[19] Our modeling admits such an alternative interpretation, with the added implication that it does not matter whether or not a design is double-blind.

A related intriguing reflection is that if DM is his own audience, in this sense, that may strengthen the justifications for assuming that there is complete information about θ, and that DM would play a SE. Typically, in game theory, justifying coordination on some particular equilibrium is a non-trivial proposition. However, if DM is his own audience, we have a one-player game, so forming equilibrium expectations should be easy.[20]

[19] 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.

[20] One may wonder if, in the spirit of Kőszegi's (2010) "preferred personal equilibrium," it would then matter if one imposed that DM would self-select a SE that gave him the highest feasible utility. If one restricts attention to SE with full-support-on-y the answer is no, since all those give 0-utility (see section 3.3). And, in fact, if θ is high enough DM must get a utility of 0 also in any other SE, as one can infer by reasoning similar to what we did in the paragraph preceding Example 4.

Self-concept maintenance MA&A propose that people are often torn between two competing motivations: gaining materially from cheating and maintaining a positive self-concept as honest. People solve this dilemma by striking a balance between the two objectives. MA&A call this "self-concept maintenance," which F&FH suggest is relevant for explaining their findings.

What does it mean to maintain a self-concept as honest? The notion has many facets. One aspect may be to internalize a preference 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 of cheating. 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 MA&A's ideas.

References

Abe, N., Greene, J., 2014. Response to anticipated reward in the nucleus accumbens predicts behavior in an independent test of honesty. J. Neurosci. 34, 10564–10572.

Abeler, J., Becker, A., Falk, A., 2014. Representative evidence on lying costs. J. Public Econ. 113, 96–104.

Abeler, J., Nosenzo, D., Raymond, C., 2016. Preferences for Truth-Telling. Unpublished manuscript.

Akerlof, G., 1983. Loyalty filters. Am. Econ. Rev. 71, 54–63.

Andreoni, J., Bernheim, D., 2009. Social image and the 50–50 norm: a theoretical and experimental analysis of audience effects. Econometrica 77, 1607–1636.

Arbel, Y., Bar-El, R., Siniver, E., Tobol, Y., 2014. Roll a die and tell a lie – what affects honesty? J. Econ. Behav. Organ. 107, 153–172.

Battigalli, P., Dufwenberg, M., 2007. Guilt in games. Am. Econ. Rev. Pap. Proc. 97, 170–176.

Battigalli, P., Dufwenberg, M., 2009. Dynamic psychological games. J. Econ. Theory 144, 1–35.

Bernheim, D., 1994. A theory of conformity. J. Polit. Econ. 102, 841–877.

Blume, A., Lai, E., Lim, W., 2016. Eliciting private Information with Noise: The Case of Randomized Response. Unpublished manuscript.

Bucciol, A., Piovesan, M., 2011. Luck or cheating? A field experiment on honesty with children. J. Econ. Psychol. 32, 73–78.

Cohn, A., Fehr, E., Maréchal, M., 2014. Business culture and dishonesty in the banking industry. Nature 516, 86–89.

Cohn, A., Maréchal, M., Noll, T., 2015. Bad boys: how criminal identity salience affects rule violation. Rev. Econ. Stud. 82, 1289–1308.

Conrads, J., Irlenbusch, B., Rilke, R.M., Walkowitz, G., 2013. Lying and team incentives. J. Econ. Psychol. 34, 1–7.

Conrads, J., Lotz, S., 2015. The effect of communication channels on dishonest behavior. J. Behav. Exp. Econ. 58, 88–93.

Dai, Z., Galeotti, F., Villeval, M.-C., 2017. Cheating in the lab predicts fraud in the field. An experiment in public transportations. Manag. Sci. https://doi.org/10.1287/mnsc.2016.2616.

Dieckmann, A., Fischbacher, U., Grimm, V., Unfried, M., Utikal, V., Valmasoni, L., 2015. Trust and Beliefs among Europeans: Cross-Country Evidence on Perceptions and Behavior. Unpublished manuscript.

Diekmann, A., Przepiorka, W., Rauhut, H., 2015. Lifting the veil of ignorance: an experiment on the contagiousness of norm violations. Ration. Soc. 27, 309–333.

Dufwenberg, M., Lundholm, M., 2001. Social norms and moral hazard. Econ. J. 111, 506–525.

Ellingsen, T., Johannesson, M., 2008. Pride and prejudice: the human side of incentive theory. Am. Econ. Rev. 98, 990–1008.

Fischbacher, U., Föllmi-Heusi, F., 2013. Lies in disguise – an experimental study on cheating. J. Eur. Econ. Assoc. 11, 525–547.

Fosgaard, T., Hansen, L., Piovesan, M., 2013. Separating will from grace: an experiment on conformity and awareness in cheating. J. Econ. Behav. Organ. 93, 279–284.

Gächter, S., Schulz, J., 2016. Intrinsic honesty and the prevalence of rule violations across societies. Nature 531, 496–499.

Garbarino, E., Slonim, R., Villeval, M.-C., 2016. Loss Aversion and Lying Behavior: Theory, Estimation and Empirical Evidence. Unpublished manuscript.

Geanakoplos, J., Pearce, D., Stacchetti, E., 1989. Psychological games and sequential rationality. Games Econ. Behav. 1, 60–80.

Gneezy, U., Kajackaite, A., Sobel, J., 2018. Lying aversion and the size of the lie. Am. Econ. Rev. 108, 419–453.

Greene, J., Paxton, J., 2009. Patterns of neural activity associated with honest and dishonest moral decisions. Proc. Natl. Acad. Sci. 106, 12506–12511.

Houser, D., List, J.A., Piovesan, M., Samek, A.S., Winter, J., 2016. On the origins of dishonesty: from parents to children. Eur. Econ. Rev. 82, 242–254.

Houser, D., Vetter, S., Winter, Joachim K., 2012. Fairness and cheating. Eur. Econ. Rev. 56, 1645–1655.

Kajackaite, A., Gneezy, U., 2017. Incentives and cheating. Games Econ. Behav. 102, 433–444.

Kartik, N., 2009. Strategic communication with lying costs. Rev. Econ. Stud. 76, 1359–1395.

Khalmetski, K., Sliwka, D., 2017. Disguising Lies – Image Concerns and Partial Lying in Cheating Games. University of Cologne. Unpublished manuscript.

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

Kőszegi, B., 2010. Utility from anticipation and personal equilibrium. Econ. Theory 44, 415–444.

Kreps, D., Wilson, R., 1982. Sequential equilibrium. Econometrica 50, 863–894.

Kroher, M., Wolbring, T., 2015. Social control, social learning, and cheating: evidence from lab and online experiments on dishonesty. Soc. Sci. Res. 53, 311–324.

Luttmer, E., Singhal, M., 2014. Tax morale. J. Econ. Perspect. 28, 149–168.

Mazar, N., Amir, O., Ariely, D., 2008. The dishonesty of honest people: a theory of self-concept maintenance. J. Mark. Res. 45, 633–644.

Muehlheusser, Roider, G.A., Wallmeier, N., 2015. Gender differences in honesty: groups versus individuals. Econ. Lett. 128, 25–29.

Olken, B., 2015. Promises and perils of pre-analysis plans. J. Econ. Perspect. 29, 61–80.

Pascual-Ezama, D., Fosgaard, T., Cardenas, J., Kujal, P., Veszteg, R., de Liaño, B., Gunia, B., Weichselbaumer, D., Hilken, K., Antinyan, A., Delnoij, J., Proestakis, A., Tira, M., Pratomo, Y., Jaber-López, T., Brañas-Garza, P., 2014. Context-dependent cheating: experimental evidence from 16 countries. J. Econ. Behav. Organ. 116, 379–386.

Piff, P., Stancato, D., Côté, S., Mendoza-Denton, R., Keltner, D., 2012. Higher social class predicts increased unethical behavior. Proc. Natl. Acad. Sci. 109, 4086–4091.

Shalvi, S., Dana, J., Handgraaf, M., De Dreu, C., 2011. Justified ethicality: observing desired counterfactuals modifies ethical perceptions and behavior. Organ. Behav. Hum. Decis. Process. 115, 181–190.

Shalvi, S., Eldar, O., Bereby-Meyer, Y., 2012. Honesty requires time (and lack of justifications). Psychol. Sci. 23, 1264–1270.

Shalvi, S., Handgraaf, M., De Dreu, C., 2010. Ethical manoeuvring: why people avoid both major and minor lies. Br. J. Manag. 22, S16–S27.

Tadelis, S., 2008. The Power of Shame and the Rationality of Trust. UC Berkeley. Unpublished manuscript.

Utikal, V., Fischbacher, U., 2013. Disadvantageous lies in individual decisions. J. Econ. Behav. Organ. 85, 108–111.