WHAT ARE WE REALLY TEACHING THEM? TESTING THE SOCIALIZATION OF A MAXIMIZING MINDSET

Eli Spiegelman

Univ. Bourgogne Franche-Comte, Burgundy School of Business
29 rue Sambin, Dijon 21000, France
+33(0) 380 725 900
eli.spiegelman@bsb-education.com

16 April, 2019

Abstract: An extensive literature suggests a certain "economist effect": students of business and economics behave more like *Homo economicus* than do those form other disciplines. A central question turns on whether this represents selection of "selfish maximizers" into the disciplines, or whether studying them affects preferences. We present an experiment building on a more recent literature suggesting this dichotomy might be too simplistic. Our strongest claim to contribution is to test not just how participants' behavior changes, but also the change in how they feel about this behavior. We find evidence suggesting that while basic preferences remain relatively constant over time, students (1) learn to attach normative weight to maximizing behavior, and (2) develop some "foresight" or sophistication in how to put this behavior into practice.

JEL codes: A20; D91; C9

Keywords: Experiment; Lying; Moral Balancing; Economics and business students

I. INTRODUCTION

This paper adds to the literature showing that students who major in business and economics (henceforth "economists") behave differently from, and in particular more in line with the "ideals" of *Homo economicus*, than do students of other disciplines. Not to put too fine a point on it, they are more selfish income-maximizers. The result has been robustly demonstrated, beginning with several pioneering papers showing that economists have a greater propensity for free riding (Marwell & Ames, 1981), expect and provide less generosity in ultimatum games (Carter & Irons 1990) and engage in less charitable giving outside the lab (Frank, Gilovich & Regan 1993) than do their non-economist counterparts. With notable exceptions, subsequent literature has confirmed the effect over many dimensions of prosocial behavior. Thus, for the purposes of this paper, we presume the phenomenon to exist. The question here is more closely related to why than if it occurs.

In particular, we propose an experimental test that we argue can measure the process of *socialization of the maximizing mindset*. This is related to another stream from the selfish-economist literature, the question of selection versus "indoctrination": If economists are different, is that because of the people who select into the program, or is it caused by the study? The evidence is mixed. Carter & Irons (1991) found early that even freshman economists are more selfish than others, and that the gap does not change with time, and the evidence since then has also largely tended to confirm selection. Taking economics courses seems to make nonmajors behave more selfishly, but selection into the field explains most of the behavior of economist themselves. (e.g., Frank & Schultze 2000; Frey & Meier, 2003, 2005; Rubinstein, 2006; Bauman & Rose 2011, van Andel, Tybur & Van Lange, 2016; López-Pérez and Spiegelman, forthcoming, Haucap & Müller, 2014; Wang, Zhong, & Murnighan, 2014; Wang, Malhotra & Murningham, 2011).

However, the issue remains stubborn. As Frank, Gilovich & Regan (1981, p. 170) put the point, "it would be remarkable indeed if none of the observed differences in behavior were the result of repeated and intensive exposure to a model whose unequivocal prediction is that people will defect whenever self-interest dictates." Kirchgassner (2005) points out further that it must be taken to some extent as a failure of learning if our education in the "economic way of thinking" *doesn't* have any effect on our students.

A subtlety of this literature turns on different levels at which behavior may be determined. On the one hand, the idea implicit in the characterization of economists as closer to *Homo economicus*, for instance in the opening sentence of this paper, calls upon the notion of preferences. The suggestion is of differences across disciplines in basic moral reasoning, attitudinal positions reflecting deep beliefs about what is important. Economists and others simply disagree on what is "bad". On the other hand, social behavior is influenced not only by preferences but also by beliefs, often in the form of social norms which may be "layered over" the basic preferences (Ostrom, 2000). For instance, the term *conditional cooperation* has been coined to refer to people who have a "preference" for following the expected behavior of others with whom they interact – a local social norm. Psychologists further distinguish between social and personal norms, where the latter are internalized, their violation generating feelings of shame and guilt, while the former are "enforced" mainly through expectation of social disapproval.

The possibility that the "economist effect" is actually an issue of norms has been raised before (Goshal, 2005; Etzioni 2015; Caviola & Faulmüller, 2014). Economists may have basic preferences no different from those of other people, and yet behave differently because they have different expectations about others' behavior. That is, they may agree on what is "bad", and still diverge on what is "justified in the circumstances." Notice that by this reckoning, social norms will only be behaviorally relevant in cases where the norm dictates behavior different from what would come from underlying preferences. Therefore, conforming with a social norm implies acting against one's own preferences, somehow. Studying this requires teasing apart what people do from what they "truly believe". One methodology for addressing the problem has used survey evidence. Researchers looking for differences in underlying preferences ask about the perceptions of fairness of market allocations (Kahneman, Knetsch & Thaler, 1986; Frey, Pommerehne, & Gygi 1993; Cipriani, Lubian, & Zago, 2009; Cappelen, Sørensen, & Tungodden, 2010; Gerlach 2017), moral judgment tasks (Hummel, Pfaff & Rost, 2016; Racko, Strauss & Burchell, 2017), political ideology (Delis, Hasan & Iosifidi, 2017) or psychological personality type (Krick, et al., 2016). This literature suggests that indeed, the more basic, "low level" traits such as moral competence and political ideology, do not vary dramatically either over economics study, or between economists and others. However, "higher level," more socially constructed phenomena, and particularly expectations of others' behavior, are more clearly related to discipline. For instance, in a recent study, Gerlach (2017) finds that differences in generosity between economists and others, while pronounced, are not reflected in different ideas of fairness: economists and others offered similar (non-incentivized) notions of what a fair offer to divide a sum of money would be. However, the differences in generosity do covary with expectations about what others will offer, which are well predicted by discipline.

These survey and questionnaire data suffer from both hypothetical and social desirability biases. Even if participants are not influenced by what they think the experimenter wants to hear, it may be difficult to sift through one's own thinking about what true moral principles are, versus what can be justified by social or contextual specifics. The issue therefore merits more study. In this line, one of the main contributions of the current study is that it provides a behavioral measure of moral intuition: an incentivized empirical test of the change in the moral weight of a particular act – in this case, an efficiency-improving lie – on participants before and after completing the major part of their coursework in a Masters in Business program at a business school in France. We therefore improve the state of the literature not just by going beyond self reports of moral reasoning. We also begin to distinguish between the categories of "bad but justified" social norms, and "justified" personal norms. The vision of the economist effect we study is one in which social norms of maximizing behavior become internalized over time as personal norms. To do this, we harness another phenomenon of the empirical literature, the moral licensing/moral cleansing effect, or moral balancing (Ploner & Regner, 2013; Mazar & Zhong, 2010; Saschdeva, Iliev & Medin 2009; Nisan & Horenczyk, 1990; Merritt, Effron & Monin, 2010; Clot, Grolleau & Ibanez, 2017; Seçilmis, 2018). Closely related to the idea of cognitive dissonance reduction (Festinger, 1957; Beauvois & Joule, 1996), this phenomenon holds that people try to maintain a "good enough" moral self image by balancing good and bad acts. It has been shown that, on the one hand, a "good" moral self-image sets up a sort of moral credit, giving people some leeway to commit later "misdeeds", while a "bad" act can be paid for in a way with later "good" behavior. Note the implicit assumption, however, which is that the one deed is considered by the decision maker to be "bad" while the other is considered to be "good". For individuals who do not think of the first as "bad", moral licensing predicts less of the subsequent (or preparatory) "good". Therefore, the extent of compensatory behavior gives an observable, incentive-compatible measure of just how "bad" the individual in question felt the previous state to be.

Specifically, we compare a treatment in which participants earn money through dishonesty with a treatment in which the same monetary endowments are distributed exogenously. The lie that maximizes profits in the former is the “bad” behavior; the latter implements the same outcomes without running down moral credit. Participants then have an opportunity to share the gains with another player – this is the costly “good” behavior that moral cleansing predicts as a response to the moral self-image that has been threatened by the lie. These two treatments are repeated for both first-year and third-year students to measure the effect of the program.

The socialization hypothesis suggests that social norms among economists, based on expectations of others’ behavior, become internalized personal norms through participation in the program. If this internalization reduces the moral weight of the lie, then the prediction would be that that moral licensing and/or cleansing effects should diminish over the course of the business degree program. Pure selection effects, ceteris paribus, would predict no change over time in moral licensing. To the extent that “bad but justified” still retains a core of “bad”, we can distinguish between “adoption of norms” that run counter to preferences, and “internalization of norms”, that is, between “bad but justified” and “justified”. The table below summarizes the general direction of our hypothesis.

TABLE 1

		Third year	
	Moral cleansing	Occurs	Does not occur
First year	Occurs	Pure selection in social norms. Behavior is driven by a “culture of non-cooperation” that goes counter to underlying type	Socialization effect. Personal and social norms are originally in conflict, then converge
	Does not occur	Resistance, then succumbing to social norms? (Not expected.)	Pure selection in personal norms. Behavior is driven by underlying type

Interpretation of the patterns of moral cleansing over the course of studies.

To summarize, the principal question we seek to address concerns the change in the moral weight of lying across years. Specifically *Is the effect of dishonesty on sharing constant over time?* Norm internalization implies that lying itself should remain essentially constant across years, but that the difference in compensatory sharing across the lie and no-lie treatment conditions should diminish. Pure selection, by contrast, predicts a constant relationship between the source of the money (lie or no-lie) and the subsequent transfers. Note that this cannot be addressed by the analysis of the dishonesty on its own, since that is, by this reckoning, possibly constant.

On the other hand, we can also address two subsidiary questions that aim to test the change in basic "behavioral preferences" with comparisons of upper- and lower-classmen in the lines of the literature beginning with Carter and Irons (1993). First, *Does the extent of lying change over time?* This represents behavioral preferences in the sense that the lying in this experiment is co-determined with the subsequent sharing decision, and so any change would not identify a change in the absolute (moral) cost of lying, but something more along the lines of the social norm. The second subsidiary question is, *Without dishonesty, do first- and third-year students transfer different proportions of their endowments?* Again, there may be many reasons for sharing, and this comparison cannot disentangle all of them. However, by restricting the comparison to conditions without dishonesty, we can at least rule out the effect of the lying itself.

Briefly, the answers we find to these questions are "Change", "Change" and "No change". We find a strong moral cleansing effect in the first-year students that evaporates entirely by third year. Regardless of the reason, it seems that third-year students feel no need to compensate their dishonesty with higher transfers. Further, the difference does not seem to come from changes in basic preferences. When getting the money does not require dishonesty, transfers are the same in the first and third year, controlling for endowment. The most surprising result is that the extent of lies *decreases* slightly over time: third-year students lie significantly less than first-year students. This fact, which was not particularly expected, opens an intriguing possibility. By third year, students have become *sophisticated liars*, reducing their level of dishonesty enough (below the first-year baseline) to justify a lower level of costly "moral repayment" later. This interpretation is bolstered by the finding that third-year students end up earning more money than first-year in the condition where lying is possible.

The basic socialization hypothesis does not predict this change in sophistication, so we introduce a theoretical model to organize those results. The model contains three basic components: a *moral credit* that is potentially run down by choices; a set of *personal norms* that act as a threshold for moral costs to be incurred; and two *types* of decision-maker, differing essentially in future cost discount rates. The results explain the data quite well, but only if we allow both norms and type distributions to change across years. In first year, students are (largely) *naïve liars*, violating their personal norms and engaging in moral cleansing when possible. Then they become sophisticated, resulting in a different (and specifically more lucrative, if not "optimal") pattern of lying and cleansing. If this pattern of behavior is also internalized as a personal norm, then by third year, it no longer generates the moral dissonance requiring cleansing. This story has a *post hoc* nature, and we should be wary of lending too much weight to it without further research. It also lacks a mechanism by which norms and type might change. But it does fit quite precisely the story of socialization of a maximizing mindset at the base of many intuitions of the economist effect, and falls neatly along the lines of the data.

In the following sections, we describe the experimental design, then present the data and results. We then present our simple theoretical model of moral balancing and compare it to the stylized empirical facts from the experiment. We finish with some conclusions and directions for future work.

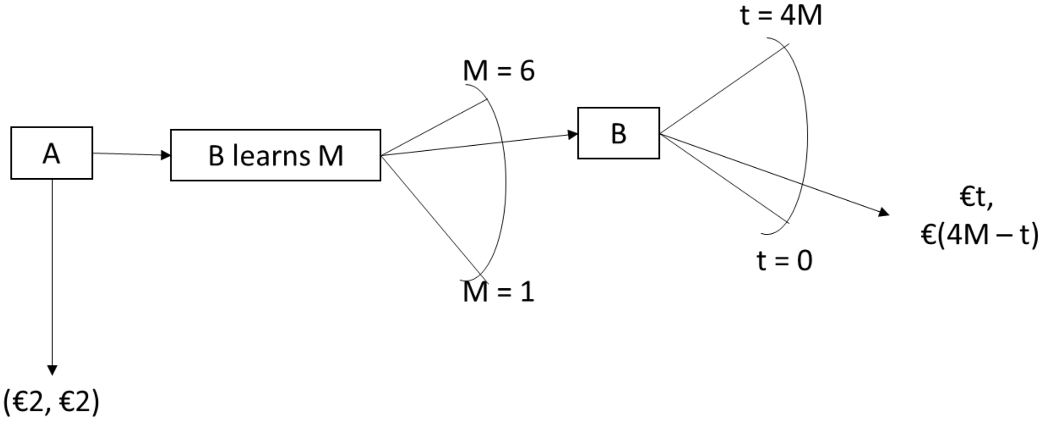
II. EMPIRICAL DESIGN

Design context: The experiment we report embeds a die-under-cup paradigm (Fischbacher & Follmi-Heusi, 2013) into a trust game (Berg, Dickhaut & McCabe 1995). The latter is a staple design of experimental economics to measure trust and reciprocity. Johnson & Mislin (2011) provide a meta-analysis of 162 replications of the game, with more than 23,000 total participants. In the canonical version of this two-player game, player A, the trustor, is given an endowment of \$10, and can transfer any part s ($0 \leq s \leq 10$) of that to player B, the trustee. The amount transferred is multiplied by 3, and the trustee can then transfer any amount t ($0 \leq t \leq 3s$) back to the trustor. Although there was significant variation in the studies, the average amount sent overall was roughly half of the initial endowment, and the average return rate was 0.37 of the (multiplied) amount received. A relevant finding for the design of the current paper was that changing the multiplier does not affect the level of trust, but increases return rates less than proportionally.

The key design innovation in our study varies this multiplier parameter, based on the results of the die-under-cup experiment. The canonical die-under-cup design gives participants a six-sided die, which they roll in private. The payoff they get depends on their report of the result of the roll. It is a useful design because it gives participants a double-blind method to engage in beneficial dishonest behavior, which can still be detected at the aggregate level. Participants can lie with true impunity, as their individual deception cannot be determined even by the experimenter. However, the distribution of declared rolls can easily be compared to the theoretical uniform distribution predicted by honest reporting. And under the assumption that people are more likely to bias their reports towards higher-payoff values, the declared roll can be taken as a (noisy) measure of dishonesty. This, combined with the simplicity of implementation, have made this design another standard feature of many experiments. It is, for instance, one of the main categories covered in several recent meta-analyses of deception games. Rosenbaum, Billinger & Stieglitz (2014 p. 186) call it the "most prominent subclass of honesty experiments", and Abeler, Nosenzo & Raymond (2016) focus entirely on designs of this type, including 72 studies and 362 individual treatments, with more than 32,000 participants. It is also one of the main subclasses in a recent study of 470 deception experiments Gerlach, Teodorescu & Hertwig (2019). Naturally, most of these studies focused on different questions, but it is worth noting in passing that the "economist effect" on lies has been part of the research programme. Lewis et al (2012) find that economists lie more, while Houser, Vetter and Winter (2012) report results among the exceptions to the "economist rule" in a coin-flip version of the experiment – economists in their study reported the payoff-maximizing result less often than did other disciplines. The central phenomenon in the literature, though, is that of "partial lying". People do lie in the design, but not to the fullest extent possible. The obvious money-maximizing choice is to report the roll that yields the highest money payoff, regardless of what was actually observed. However, on a six-sided die, the modal response is often, for instance, five when the payoff-maximizing response is six (Abeler, Nosenzo & Raymond, 2016).

Implementation: In our design, summarized in Figure 1, below, player A begins with €4, which he can either split evenly between himself and B, resulting in the end of the game with each player earning €2, or pass to B. In this second case, the €4 is multiplied by a number called the multiplier (*M*) between 1 and 6, giving B some amount between €4 and €24, of which any amount (*t*) can be transferred back to A.

FIGURE 1



ıstructions

i ısly by

ıe €4 (*No-*

t ment of 4*M*,

and the amount *t* to be passed back to A. The instructions reminded B that this transfer would only be implemented if A did in fact Trust. This is a partial strategy method because it gives information about B's choice for all possible actions by A, but not for all possible values of *M*. We discuss the implications of a full strategy method in the conclusion. Instructions used neutral language for the actions. Then the sheets were collected and matched together to generate the result of the interaction. Subjects were paid by a research assistant several days later.

The experiment used a 2x2 between-subjects design, varying first the source of *M*, and then the population on which the study was conducted. Regarding the first, in some (*Roll*) sessions, the multiplier was the report of a six-sided die, rolled privately. Visually-isolated participants in the B role were asked to roll a fair die provided to them and write the value on the experimental instruction sheet. They were encouraged to roll several times before making their final roll that "counted". The reported rolls from these sessions were then used to create a distribution of values, which were randomly allocated to further (*No-Roll*) sessions by transcribing them to the experimental instructions. Participants in these sessions were told that the number had been drawn from a particular distribution that could take any value between 1 and 6, but was skewed towards high numbers. The result of this procedure is that aggregate multiplier values were controlled across these treatments, and the only difference was the source of the number.

We also elicited Bs' expectation of trust by A in a non-incentivized manner, asking them to fill in a Likert scale indicating their beliefs about the likelihood that A would not end the game. This number ran from 1 ("Very unlikely") to 7 ("Very likely").[1] We take this to be a proxy for B's second-order expectations about A's payoff, that is, B's belief about A's belief about B's behavior. It would have been feasible, and in some ways useful, to ask this expectation directly, but in our opinion not necessarily more valid. First, it is a difficult value to assess. A's expectation is a compound average, over all the possible values of M, of the expected return conditional on that value being realized. This means B requires second-order expectations of both A's beliefs about *M* and A's beliefs about *t* given *M*. Denoting B's possible beliefs about *t* and *M* with upper and lower indexes, A's second-order belief about B's expectation is formally given as

$$E_2^B = \iint \sum_{m=1}^{6} \sum_{s=0}^{4M} s\, Pr_B^t(t = s \vee M) Pr_B^M(M = m) dF(Pr_B^t) dF(Pr_B^M)$$

[1] A's expectations about the amount that would be returned for any realized value of *M* were also elicited, providing a bonus of €2 to A if the expectation was correct. However, we do not focus on these values. First, they are subject to a hedging problem, as A players can reduce the expected variance of their earnings by declaring a value opposite to their true beliefs. Second, and more importantly, our central interest was on B's behavior, and so the experiment was implemented with most players in the B-role, and we have relatively few observations of A. While this arguably comes close to deception, the instructions indicated that the payoff of each A was determined by the choice of one B, while that of each B was determined by the choice of one A, which is true. We note this method is not uncommon in the experimental literature. See for instance Houser and Schunk (2009).

This calculation is highly subjective, given that none of the relevant distributions comes with explicit probabilities; all are more “horse race” that “roulette wheel”. The resulting ambiguity will probably add noise to the estimate, and quite possibly self-serving bias. Previous literature (Felson 1981; Dunning, Meyerowitz et al. 1989; Dana, Weber et al. 2007; Haisley and Weber 2008; Valdesolo and DeSteno 2008; Sloman, Fernbach et al. 2010; Feiler, 2014) has suggested that ambiguity and vagueness enhance the capacity for self-deceptive distortion of beliefs. Given all of this, it seems plausible that B either does not have explicit access to, or simply does not formulate even implicitly, precise numerical values for A’s expected return. And while a number that B determines *post hoc* to describe the second-order beliefs may be correlated with B’s general thinking about the decision problem, the fact that B can determine a number obviously does not mean that number is causally constitutive of the previous decision. But it may tempt us with a kind of fallacy of misplaced concreteness (Whitehead, 1997). On the other hand, the choice by A to continue the game or not is directly payoff-relevant to B, and the question permits a vaguely phrased answer to match the vague expectations that are likely to exist, psychologically speaking, before the question is posed. Finally, the pen-and-paper implementation of the design makes it logistically difficult to separate the answers to different questions, and eliciting expectations over A’s behavior seemed less prone to inducing demand effects than did those about how much B thought A expected to win.

This basic experimental design was repeated twice, once with students in their first semester (*Year* = *Y1*) of a Masters in Business program, and once with students in their third year (*Y3*), just before they left to do a professional internship program. The students in this program are a highly selected group, and demographic variation across years is limited. Attrition from the program is on the order of five students per year, out of a full student body of roughly 2500.[2]

Discussion of the design: Our design is relatively similar to that in Ploner & Regner (2013). Like that paper, in this study we give participants an opportunity to pay for dishonest reporting of a private random event through subsequent generosity. A minor difference is that those authors implemented a binary result, where odd rolls (i.e., 1, 3 or 5) generated a high endowment for the trustee, while even (2, 4 or 6) rolls generated a low endowment. To allow for the possibility of partial lying, we preferred the finer-grained measure of dishonesty that comes from six possible declarations. More important differences, related to somewhat different experimental questions, included the addition of other treatments before their experiment, designed to manipulate the moral credit that participants had at the beginning, and a control (*Bonus*) treatment without division of the endowment at the end. One potential strength of their design compared to ours is that instead of *Roll* and *No-Roll* treatments, they have *Hidden* versus *Open* rolls, where the latter were observed by the experimenter to rule out cheating. Therefore, if the act of rolling itself influences behavior, then that effect will be confounded with the cheating effect in the current study.

[2] However, we note that in second year, a substantial number of new students enroll in the program. Therefore in our Y3 group, there are perhaps 20% who did not go through the Y1 year. This means that (a) the groups are not exactly the same, and (b) the number of years of training in the program is not even across the sample. Unfortunately, we did not collect data on how many years each student had been in the program. Regarding the first, in terms of basic demographics the influx is not substantially different in each year, but there is a clear concern about selection issues. Regarding the second consideration, a shorter period for some subjects might reduce any "school effect", but should lead only to false negatives in this kind of study.

Several other aspects of our design may have encouraged relatively high declarations. For instance, the practice rolls we encouraged from participants were ostensibly designed to verify that the die was indeed fair. It also has been shown, however (Shalvi et al. 2011) that observing counterfactual high rolls encourages participants to report high values. Second, Ploner and Regner (2013) compare the baseline Bonus treatment to ones where the endowment is split, finding what they call a “Robin Hood effect” in which division of the gains encouraged dishonesty. These factors, as well as the fact that out subjects were from a business school, should be taken into account when comparing the aggregate honesty results we obtain to those in the literature.

It is also important to emphasize that the central non-experimental variable in this design – participation in the business school program – is measured through the proxy of the number of years’ study. Therefore, we must admit the possibility that other variables may correlate with this measure and confound the results. For instance, Y3 participants are, logically, older than Y1, so we cannot rule out that any effect comes simply from aging. On the other hand, the two-year time lag is not long; the meta-analysis by Gerlach Teodorescu and Hertwig (2019) suggests that on average, two years’ difference in age reduces cheating by about 3.4%. While these may arguably be formative years, if the effect is specific to the formation over those years, then that further explains the phenomenon, rather than confounding it.

Another possible confound is an end-game effect.[3] According to this idea, Y3 students have a shorter future time horizon with their classmates, and therefore a smaller potential “feasibility set” of cooperative actions, since rewards or punishments for current behavior will be smaller in cumulative value. This would then predict less cooperative behavior in Y3 than in Y1,

[3] We thank an anonymous reviewer for raising this interesting point.

confounding the "economist effect". This is possible, although of course the anonymous nature of the interaction makes it more difficult to sustain. We do not have solid evidence on this line, but anecdotally, if anything Y3 students appear to have tighter social bonds than Y1, since they have more history. It has been shown that reducing social distance encourages cooperative behavior, so this force should work against the prototypical economist effect.[4] It is worth keeping this point in mind, but overall, we would expect the "push of the past" to outweigh the "pull of the future". Finally, this cross-sectional study has the usual benefits and drawbacks of its design. In particular, we cannot control explicitly for constant individual characteristics, because the Y3 and Y1 groups are different subjects, as are the Roll and No-Roll. On the other hand, in addition to reducing the time required to run the study and avoiding the logistical strain of maintaining anonymity in a wide-spaced panel, the between-subjects design reduces learning effects specific to this experiment. It was the first time that both groups played the game, which makes their behavior arguably more comparable than if the same individuals played it twice. And because the group is relatively homogenous, many of the worst potential confounds that might arise – such as those alluded to above – are intertemporal in nature, and would not be eliminated by a panel design. Still, all the factors mentioned above should be taken into account when interpreting the results below.

Empirical hypotheses: As noted in the introduction, our key contribution is the measure of the change in the moral weight of dishonesty. For each year, we measure this by comparing the average of the transfer rates (*s*), defined as the percent of the multiplied endowment that B

[4] This is at least true with respect to the transfer decision. It is unclear what the predictions for the end-game and social distance effects would be for the lying decision. The lie in this case, because it increases the size of the total "pie", is in some was a pro-social choice.

proposed to transfer (i.e., $s = t/4M$), between the *Roll* and *No-Roll* conditions. The overall distribution of endowments is the same in both conditions; the only systematic difference between *Roll* and *No-Roll* is the source of the money. Therefore, systematic differences in the share rates across these conditions should indicate some causal relationship between the source of the money and the transfer. The causal relationship we have in mind is of course moral balancing. This predicts that the dishonesty in the *Roll* condition will cause participants to feel some dissonance, which can be reduced by transfers to A. Since this compensatory balancing is costly, we suppose that people will engage in the lowest amount sufficient to repay the moral debt (cf Grossman, 2015), so the transfer will be increasing in the moral cost incurred. This implies that, *ceteris paribus*, greater transfers will indicate greater dissonance caused by the dishonesty. As a result, the difference between the average transfer in the *Roll* and *No-Roll* conditions measures the average level of dissonance generated by the dishonest reporting. Jacobsen, Fosgaard & Pascual-Ezama (2017) report that there is evidence that lying is susceptible to this kind of moral balancing effect. The null hypothesis that holds if this moral balancing does not change over time takes the form of a difference-in-differences. Specifically, the difference in average percent transfer between the *Roll* and *No-Roll* in *Y1* should be the same as the analogous difference in *Y3*. Identifying the transfer rate as *s* and indicating the two treatment variables as *Roll* and *Year*, we test this null with the regression equation

$$s = + \quad {}_{R}Roll + \quad {}_{Y}Year + \quad {}_{X}RollYear + \{Control\}$$

In this estimation, the difference in differences is identified as a non-zero coefficient on the interaction term. In the controls, we include the multiplier value for the individual, as well as gender and the reported expectation that A would Trust.

Among the subsidiary questions identified in the introduction, one concerned distributional preferences, which are measured by comparing transfer rates across years for participants who did not lie, that is, in the *No-Roll* condition. As mentioned earlier, this behavioral difference may include several different underlying motivations, in this case for example reciprocity or guilt aversion. We discuss guilt aversion later in the results, but in either case it is important to control for expectations, which we do with the proxy variable of B's expectations of trust. The regression equation is therefore

$$s = + \quad {}_{R}exp + \quad {}_{Y}Year + \quad {}_{X}expYear + \{Control\}$$

where EXP is the reported likelihood of trust and *s* is the average transfer rate. In the controls are gender and the value of the multiplier. The interpretation of the coefficient on EXP is the effect of expectations of trust within Y1 subjects. The interaction is the difference in the effect of trust on the two years, and the coefficient on the year is the "baseline" difference in transfers across year, holding trust expectations constant.

The final question concerns the average rate of dishonesty across years, and is tested with a comparison of the distribution of reported rolls in the *Roll* condition in the two years. While indicative of differences over time, we do not claim this to be a perfect measure of changes in pure lie aversion. First, as mentioned above, a preference for efficiency is a countervailing incentive that may weigh against lie aversion. Even the lie-averse may do so when the benefit is great enough. Also, participants know when they make their report that the "rewards" to the lie will be split with player A. Barr and Michailidou (2017) report that lies are more common when the rewards benefit many people. We can investigate this factor by controlling for transfers in the comparison of declarations. Finally, this comparison does not account for the moral weight of the lie, which is what we investigate in the first question. In the best-case scenario, the comparison of lie rates measures the change in "bad but justified" behavior, that is, changes in the social norms across the years.

III. DATA AND RESULTS

Sample and data description: Our sample includes *N* = 578 observations in Role B, collected in 24 paper-and-pencil sessions between September and December, 2017. Table 1, below, shows the breakdown of the experimental treatments.

TABLE 2

Variable	*No-Roll*	*Roll*	Total
Y1	193 (7 sessions) 44.6% Female	127 (5 sessions) 40.9% Female	320 (12 Sessions) 43.1% Female
Y3	139 (6 sessions) 47.5% Female	119 (6 sessions) 45.4% Female	258 (12 Sessions) 46.5% Female
Total	332 (13 Sessions) 45.8% Male	246 (11 Sessions) 43.1% Male	578 (24 Session) 44.6% Female

Experimental cells.

The table shows that there were slightly more females in Y3 than in Y1, and slightly more in the *No-Roll* condition than in the *Roll* condition. However, none of these differences are significant (pairwise rank-sum (Mann-Whitney) test for pairwise equality of medians all over 0.31). All participants were students of [Name of school removed] in the Masters of Business program, either in the their first term (Y1) or in the first term of their third and final year (Y3). Groups go through the program in cohorts, with a number of common courses, (including courses on business ethics) in addition to a number of possible specializations. The participants in this experiment were recruited from the common courses, so were a representative sample of all the different specializations at the school. The experiment took about 45 minutes in total, and average earnings among the B-roles were around €14.02.

The table below shows the overall means and standard deviations for the main variables in the study, as well as those for each year separately, pooled over the *Roll* and *No-Roll* conditions. Looking at the totals column, we see the first evidence of dishonesty in the average multiplier rate, which is more than 5 out of 6. Recall that honest reporting of the die roll would result in a mean of about 3.5, and standard deviation of about 1.71, which puts the observed average nearly 22 standard errors away from that predicted by honest reporting. We also see that generally, B expected A to trust, giving a value of over 5 on a scale of 1 to 7 (the actual proportion of trust was 0.831). The average multiplier values imply that B, on average, had around €20 endowment of which slightly less than a third, or €6.31, was transferred back. We note this is in line with the standard results in the literature, as cited for example in the meta-analysis cited above.

TABLE 4

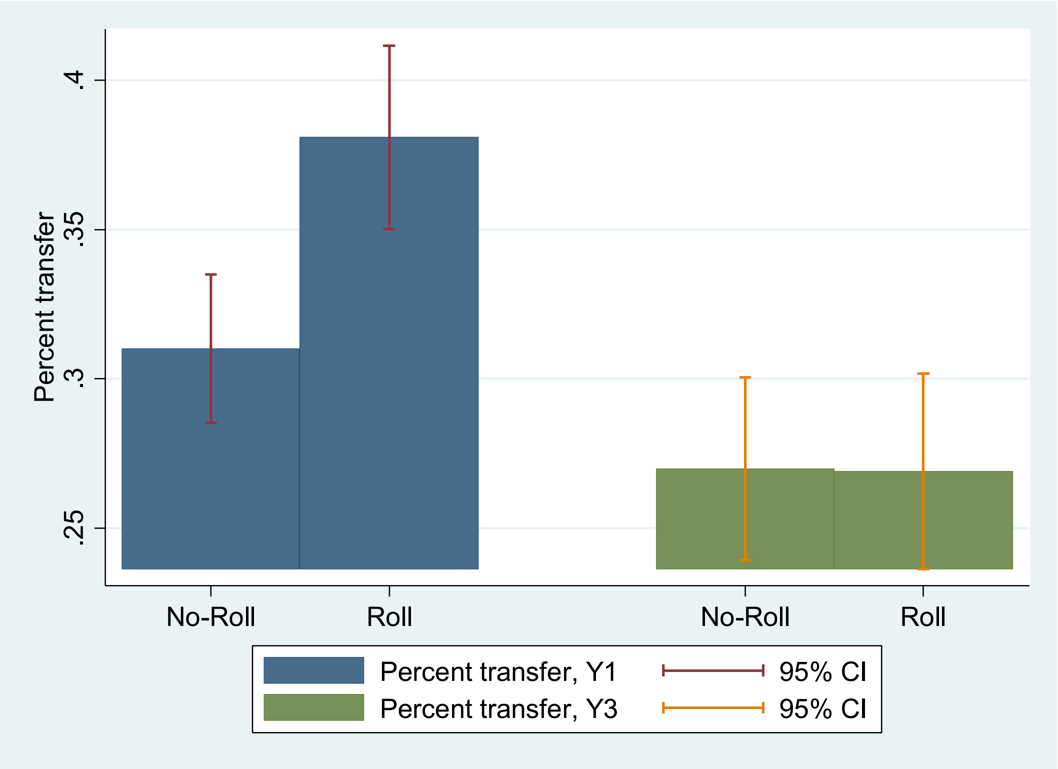
Variable	***Y1***			***Y3***			***Total***		
	mean	sd	N	mean	sd	N	mean	sd	N
Female	0.44	0.50	308	0.45	0.50	253	0.45	0.50	561
M	5.18	0.98	320	4.90	1.12	258	5.06	1.06	578
Expectation of trust	5.20	1.52	305	4.91	1.51	250	5.07	1.52	555
Amount transferred	6.94	3.90	320	5.30	3.81	258	6.21	3.94	578
Percent transferred	0.34	0.18	320	0.27	0.18	258	0.31	0.18	578
Payoff	13.78	4.73	320	14.32	4.92	258	14.02	4.82	578

Mean and standard deviation for the main variables in the study. Note that different sample sizes are due to some participants' failing to answer gender and expectation questions.

Comparing the two years, a somewhat conflicting picture emerges. Y1 participants reported higher values of M, and also higher expectations of trust. On the other hand, they transferred a substantially higher amount back – a Mann-Whitney test of equal medians for the percent transferred yields a p-value less than 0.0001. As a result, they ended up earning less money, although that difference is not significant (Mann-Whitney p = 0.1667). We investigate these differences in more detail through our hypotheses specified above.

Result 1 –moral balancing: Here we compare the difference in the percent of the endowment transferred in the Roll and No-Roll conditions in Y1, with that in Y3. The Figure below illustrates the basic result.

FIGURE 2



Average transfer rates by year and treatment condition

The clear message from this figure is that there is a strong moral balancing effect in Y1 that completely vanishes in Y3. The average transfer rises from 0.310 to 0.381 in Y1 when the source of the endowment involves dishonesty, which implies an extra €1.50 or so transfer. In Y3, the rate actually falls, albeit by an almost unmeasurable 0.00089, corresponding to less than €0.02. A Mann-Whitney test of the first difference shows a significant effect ($p < 0.001$), while the second is unsurprisingly not distinguishable from zero ($p > 0.9$).

A direct comparison across years is made more difficult by the fact seen above that the different years reported different rolls. Using the percent transfer controls for this, but it is still possible that the percent transfer depends on the overall endowment. The regressions (1) and (2) below control for the multiplier value, expectations and gender. While in principle each participant represents an independent observation, we cannot rule out interaction effects at the session level, as the participants did know each other outside the lab. We therefore opt conservatively to cluster standard errors at the session level in the regressions reported.

TABLE 4

	(1)	(2)	(3)
Outcome	Transfer	Transfer	Payoff
Y3	-0.0320	-0.0222	0.555
	(0.0276)	(0.0275)	(0.536)
Roll Condition	0.0809***	0.0744***	-1.616***
	(0.0213)	(0.0249)	(0.501)
Interaction	-0.0807**	-0.0842**	1.707**
	(0.0330)	(0.0348)	(0.692)
Male	-0.0300	-0.0228	0.598
	(0.0175)	(0.0164)	(0.369)
M	-0.0116	-0.0260**	3.054***
	(0.00905)	(0.00998)	(0.142)
Expectation of trust	0.0190***	0.0197***	-0.395***
	(0.00658)	(0.00661)	(0.140)
Constant	0.311***	0.375***	-0.224
	(0.0503)	(0.0706)	(1.017)
Observations	541	406	541
R-squared	0.094	0.090	0.440
Robust standard errors in parentheses			
*** p<0.01, ** p<0.05, * p<0.1			
(2) Limited to multiplier > 3			

Regression results for the effect of the *Roll* condition on percent transfer (regressions 1 and 2) and on the Payoff (regression 3)

In these interacted regressions, the Y3 dummy shows the difference between the years in the *No-Roll* condition (first and third bars of Figure 5, above), which is seen not to be significant, showing incidentally that changes in prosocial attitudes are not driving these results. The direct coefficient on the *Roll* condition shows the effect of the treatment on Y1 participants (first and second bars), which can be seen to be significantly positive in regressions (1) and (2), even controlling for gender, expectations, multiplier values and session-level effects. The interaction is interpreted as the difference in differences, which is of very similar magnitude to the direct effect, but opposite sign. The sum of these, therefore, interpretable as the effect of the treatment on Y3 participants (third versus fourth bars), is close to zero and not significant ($p > 0.85$ for both sums). Interestingly, expectations significantly predict transfer rates, which bolsters our interpretation of that variable as a second-order expectation.

Regression (2) repeats the same regression as (1), but limited to participants with a multiplier greater than 3, in case the differences occur in the upper tail of the multiplier distribution. The results are qualitatively the same, as they are when limiting the sample further to only multipliers of 5 or 6, as shown in the table below. When limited only to multipliers of 6, the difference in Y1 is of more marginal significance, interestingly. But it is still a more significant difference than that found in Y3.

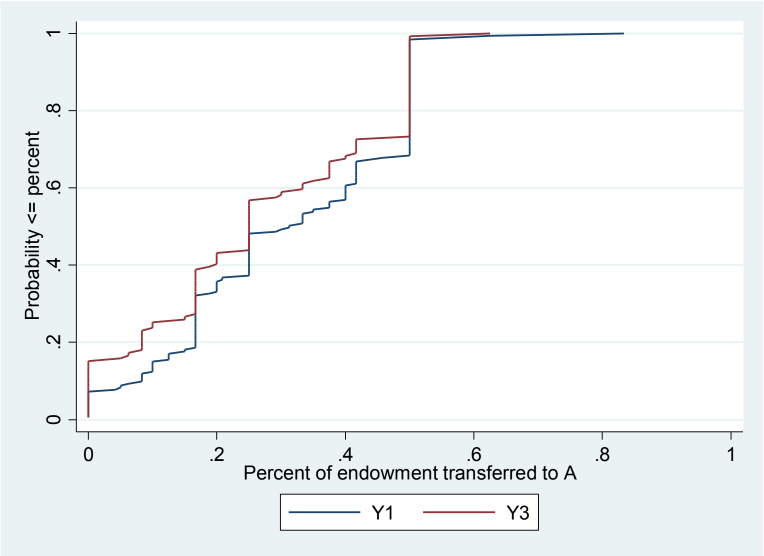
TABLE 5

Multiplier	**Year**	**No-Roll (N)**	**Roll (N)**	**Mann-Whitney p**
5	Y1	0.318 (57)	0.417 (32)	0.004
	Y3	0.258 (51)	0.284 (42)	0.682
6	Y1	0.298 (85)	0.348 (73)	0.078
	Y3	0.277 (49)	0.248 (43)	0.569

Transfer rates by year and treatment condition for M = 5 and 6

Result 2, distributional preferences. The appropriate test to compare distributional preferences across the two years compares transfer rates in the *No-Roll* condition. As for the lying effect below, the results here shouldn't be taken as "pure" distributional preferences. The fact that the transfers were contingent upon trust by A implies that they likely mix preferences over distribution with reciprocity concerns. However, controlling for expectations via our proxy allows us to address a more nuanced question, teasing apart the two effects to some degree. To begin, the figure below shows the cumulative distribution of the percent of endowment transferred in the two years in the *No-Roll* condition. Among the salient facts are (1) nearly all participants transferred less than half of their endowment; (2) there are large spikes in the distribution at 0 and 0.5, as well as around halfway through the distribution; (3) transfers were somewhat higher in Y1 than in Y3. On the last point, the average transfer proportion in Y1 was 0.310, while in Y3 it was 0.270. A Mann-Whitney test confirms that the median transfer was marginally higher in Y1 (p < 0.1), but a Kolmogorov-Smirnov test of equality of distributions does not indicate a difference (combined p-value 0.272).

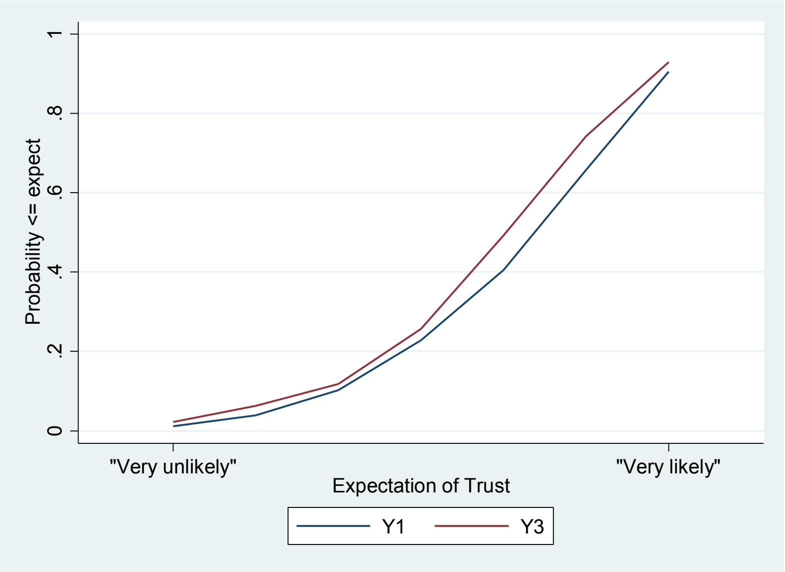
FIGURE 3



Cumulative distribution of percent transfer in *No-Roll* condition by year.

While transfers were higher, as mentioned, this mixes reciprocity with fundamental distribution. To investigate this relationship, we consider our expectations proxy. The figure below shows the cumulative distribution for these expectations, smoothed by calculating the midpoints of cumulative probability for each distinct value. Aa similar pattern emerges in this variable to that of the percent transfer above overall. Y1 participants reported a value of 5.15 out of 7 for the likelihood of trust, while Y3 participants averaged 4.875. The difference is also marginally significant with a Mann-Whitney test (p < 0.1), but not in a Kolmogorv-Smirnov (p = 0.219).

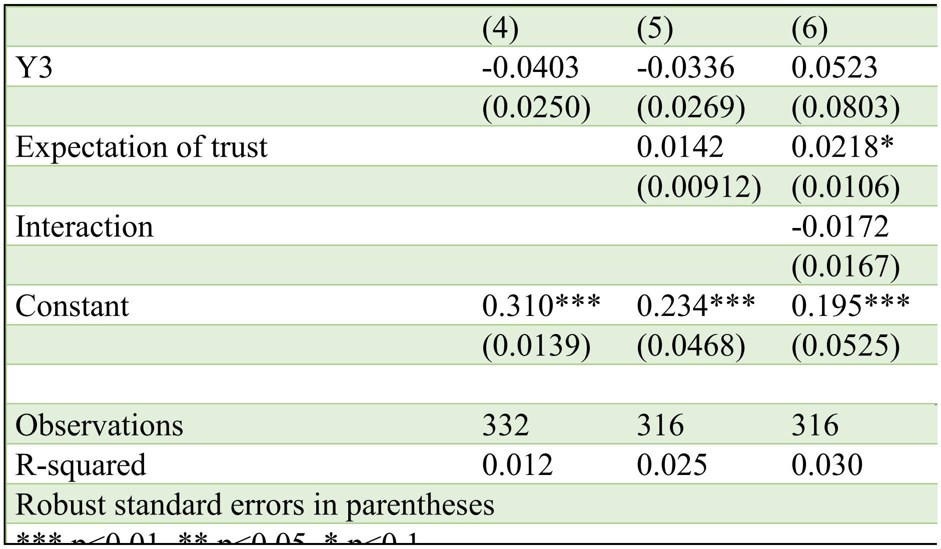
FIGURE 4



Cumulative distribution of expectation of trust, on a scale of 1 (“Very unlikely”) to 7 (“Very likely”). Distribution measured at midpoint for smoothing

In summary, Y1 participants expected more trust, and also transferred higher amounts to their A-players. Because these variables follow similar patterns individually, and in particular a relationship that would be predicted by models of reciprocity, we report several regressions on the correlation between them.

TABLE 6



	(4)	(5)	(6)
Y3	-0.0403	-0.0336	0.0523
	(0.0250)	(0.0269)	(0.0803)
Expectation of trust		0.0142	0.0218*
		(0.00912)	(0.0106)
Interaction			-0.0172
			(0.0167)
Constant	0.310***	0.234***	0.195***
	(0.0139)	(0.0468)	(0.0525)
Observations	332	316	316
R-squared	0.012	0.025	0.030
Robust standard errors in parentheses			
*** p<0.01, ** p<0.05, * p<0.1			

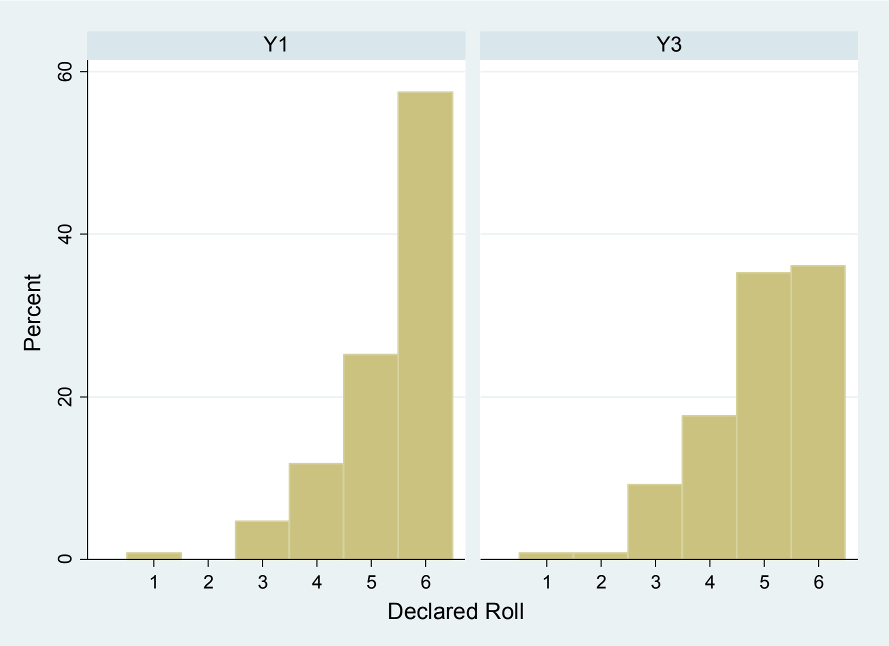
Regression results, with percent of endowment transferred as the outcome. Clustered at the session level.

Regression (4) is essentially a comparison of means, controlling for session effects. We see the marginal difference in behavior across years does not survive this control. Regression (5) adds the expectation variable as a regressor, but constrains the effect to be the same on both groups. No effect is found. In regression (6), we add an interaction term. In this specification, the coefficient on the expectation variable is interpreted as the effect of expectations on Y1 participants (i.e., when Y3 = 0), and is marginally significant. The effect of expectations on Y3 participants is calculated as the sum of the Expectations and Interaction coefficients. This sum is not significant (p > 0.5). The difference between the groups controlling for expectations is given by the Y3 coefficient, and is not significant in any of the specifications.

How should these results be interpreted? First, they once again show very little effect of studying business (at least in this school) on distributional or reciprocal preferences. Differences in reciprocity would come out in the interaction term of regression (6), and differences in basic distributional preferences should appear in the Y3 coefficient. None appear, suggesting that both of these forces are constant over time. If there is an effect, then it seems to be that reciprocity diminishes over time, since expectations have some modest effect on Y1 participants in regression (6), but none at all on Y3. Therefore, we have some weak evidence of the development of uncooperative norms in the business school, tied to diminished expectations of reciprocity. This once again lends more weight to the mechanism of selection than learning in explaining the "economist effect".

Result 3, Very high lie rates: An unexpected result finds that the overall lie rate is lower in Y3 than in Y1. The table above does not permit the correct test, because it pools over participants who rolled, and could therefore lie, and those who were given their numbers exogenously. While the distributions were the same, the sample sizes are obviously incorrect. Limiting the sample to the *Roll* condition, the average declaration in Y1 was 5.33, while in Y3 it was 4.94. Economically speaking, this is not a dramatic result. A difference of 0.4 pips would mean for instance that on average, two out of five Y3 students reduced their lie by one pip compared to the Y1 baseline. However, it is significant. A Mann-Whitney test shows that the median value is higher in Y1 than Y3 ($p < 0.001$), and a combined Kolmogorov-Smirnov test of the distributions also rejects equality ($p < 0.01$). Figure 2, below, shows the histograms of the declared rolls in each group.

FIGURE 5



Distribution of declared rolls. N = 127 (Y1); 119 (Y3)

This figure shows that although there is only a small difference in the average roll declared across the treatments, the distribution does change shape. Notably, the very marked mode of maximal lying has disappeared in Y3. Y3 participants were arguably not much more honest, because the distribution is still wildly skewed compared to the uniform predicted by honesty. Indeed, the numbers in both years are consistent with a strong selection effect of relatively non-lie-averse types into the business school. For instance, the meta-analyses of Gerlach Teodorescu & Hertwig (2017) and Abeler Nosenzo and Raymond (2016) measure the degree of lying with the statistic

$$R = \begin{cases} \dfrac{d - \acute{d}}{\acute{d} - d_{min}} & d < \acute{d} \\ \dfrac{d - \acute{d}}{d_{max} - \acute{d}} & d > \acute{d} \end{cases}$$

Where d is the declaration, $\acute{d}$ the expected declaration under honesty and d_{min} and d_{max} the minimum and maximum declarations, respectively. Gerlach, Teodorescy & Hertwig (2019) find the average value in similar experiments to be 28%, with a 95% confidence interval of 24%-32%. Applying this formula to our data, we find that the values are 57.6% and 73.2% for the Y3 and Y1 participants, respectively, both of which are far and away outside those bounds. Abeler Nosenzo and Raymond (2016) report only one individual treatment with maximal lie rates comparable to Y1 in our study.

On the other hand, there is at least no evidence that business school makes people less averse to lying. We will have a closer look to this result below, but Y3 participants show more "partial lying" (Fischbacher & Follmi-Heusi, 2013) than do the Y1. In Y1, 57.5% of subjects report a 6 on their die, while in Y3, the number is 36.1%. The latter is still significantly different from the 16.7% predicted by honest reporting (confidence interval = [27.4%, 48.9%]), but it is also significantly different from the former (Chi-square(1) p = 0.001). Naturally, without a baseline comparison among non-business students, we cannot make generalisations about the link between this effect and the business program.

Result 4 – Payoff consequences: We close with a further result that, while not among our initial hypotheses, does follow naturally from the data. We have seen above that Y1 participants lie to a greater extent than Y3 participants, and so have on average larger endowments. On the other hand, they also transfer more to A, which is obviously costly. This raises the question: which is the better strategy in payoff terms? To lie more, even if it means engaging in costly "moral compensation" later on, or to lie somewhat less, and keep more of a somewhat smaller pie? This question is addressed in regression (3) from Table 4 above. The least surprising fact from this table is that increasing the multiplier significantly increases payoff. Given the return rates upwards of 30%, it also makes sense that each pip declared, which results in an endowment increment of €4, increases payoff by about €3 on average. That the direct treatment effect – corresponding to the effect of *Roll* on Y1 – is negative also should come as no surprise, since the overall distribution of endowments was the same in *Roll* and *No-Roll*, but participants transferred more back to A in the former than in the latter. The interesting comparison is the sum of the Y3 and Interaction coefficients, which measures the difference between the years in the *Roll* condition. This is the result that allows us to compare the efficiency of the different strategies. The linear combination coefficient is 2.26, significantly different from zero ($p < 0.001$), showing that Y3 participants managed to earn significantly more money out of the *Roll* condition than did Y1.

Guilt aversion

Among the possible alternative explanations to our data, one of the most threatening appears to come from guilt aversion. We therefore will devote a small digression to its discussion here. Guilt aversion refers to the hypothesis that people are motivated in their choices to give others the payoff that they think those other people expect. It is therefore a theory that falls under the class of "psychological games" (Geanakoplos, Pearce and Stacchetti 1989; Battigalli and Dufwenberg 2007, 2009), in which beliefs directly influence utility. Specifically, guilt aversion calls on second order beliefs – B's beliefs about A's beliefs about the final payoffs, for instance – as utility arguments. This mechanism could explain the empirical results if (i) the difference across treatments in Y1 is due to guilt aversion (rather than moral balancing), and (ii) second-order beliefs change between Y1 and Y3. Such a phenomenon is plausible. In the Roll condition, it is common knowledge that B can freely decide the common income. If A expects (or more precisely if B anticipates that A will expect) B to make a high declaration and share it, then such expectations may in effect become self-fulfilling. In the No-Roll treatment, by contrast, it is common knowledge that B is bound by the exogenous multiplier, so A's expectations are harder to justify. This is indeed an alternative explanation in that it operates not through how the DM feels about her own behavior, but how she feels about how the other player feels about the outcome.

[5]For instance, consider a mechanism of *solidarity decline*.[6] Note that Y1 students (a) lie more (generate bigger pies); (b) share more; and (c) expect more trust than Y3 students. This is potentially indicative of a stronger feeling of "solidarity with the cohort" in Y1 than in Y3. The solidarity in Y1 triggers guilt aversion in the Roll but not the No-Roll treatment. Due to solidarity decline, Y3 participants in the A-Role expect lower transfers in the Roll condition; these expectations are self-realizing because B-Role participants match (rational) second-order expectations.[7]

There are, however, several reasons to doubt this explanation. First, the data on the empirical importance of second-order expectations is mixed. Charness and Dufwenberg (2006) find support, as does Khalmetski (2016), while Ellingsen et al. (2010) find less, and Vanberg (2008)

[5] In principle, a change in guilt aversion could explain a change in behavior over the two years through two classes of mechanism, corresponding to a change in beliefs or in preferences.

[6] We again thank an anonymous reviewer for this idea and term.

[7] Another interpretation of solidarity decline, moreover, could result in lower feelings of guilt for a given level of (perceived) betrayal. This would work through a change in guilt preferences. This is hard to cleanly disentangle with the data, because the moral disengagement it implies is not particularly at odds with the process of reduced moral damage of lies, and therefore does not yield dramatically different predictions. Indeed, it seems plausible that those who, as per the suggestion in the paper, feel less moral weight of lying, might also feel less guilt at keeping the resulting gains for themselves.

suggests an alternative explanation for the effect in Charness and Dufwenberg (2006). Further, Khalmetski (2016) suggests the effect might be sensitive to rationalizability of beliefs, Bellemare Sebald and Suetens find that the effect depends on the elicitation method, and Balafoutas and Sutter (2017) find that it depends on communication. It therefore appears that there are subtleties to the way in which they operate behaviorally. A plausible interpretation of the Vanberg (2008) results, broadly consistent with the subsequent literature, is that the guilt aversion effect depends on responsibility for the beliefs in question. People feel obliged to fulfill others' expectations *when they are causally implicated in generating them*. This "guilt from responsibility" argument has some philosophical plausibility, and also a potential psychological mechanism. Prior actions on B's part that determine A's expectations (such as a promises), at the same time make those expectations more salient in B's subsequent deliberation, increasing their weight in the decision. In the current experiment, there was no such responsibility, so the prior literature does not unequivocally suggest that second-order beliefs should be relevant.

More anecdotally, as mentioned above with respect to the end-game effect, while the basic pattern of the data does fit the story of solidarity decline, the cultural tenor of the school does not. The experiments were run in a small school where group projects are the norm and associative activity is particularly encouraged. The Twitter hashtag #WeAreTeamBuilding is common in school communications, and while naturally this does not preclude the existence of solidarity decline, Y3 students seem if anything like a more cohesive group than Y1.

Finally, while for the reasons mentioned earlier, we did not elicit second-order beliefs explicitly, the data we have do not fully support a guilt aversion story. For instance, we can use elicited beliefs to estimate A's first-order expectations. These are shown in the table below, and we find that they do not vary systematically or significantly either across years or across treatments. None of the differences are significant at conventional levels.

TABLE 7

	No-Roll	Roll	Total
Y1	0.42	0.37	0.40
	(17)	(12)	(29)
Y3	0.32	0.42	0.36
	(14)	(10)	(24)
Total	0.38	0.39	0.38
	(31)	(22)	(53)

Computed average return (N) expected by A

Formally, the argument turns not on A's first-order, but on B's second-order expectations, so the above comparisons are not quite correct. However, under a story of solidarity decline one would expect the two to be correlated, if not the same. And using the proxy of expectations of trust, we find no difference across the Roll (5.27) and No-Roll (5.15) treatments in Y1 (t-test p = 0.4875), which speaks against the mechanism driving that result. The table below shows the tests across all four experimental cells.

TABLE

	No-Roll	Roll	Total	t-test (p)
Y1	5.15	5.27	5.20	0.4875
	181	125	306	
Y3	4.87	4.92	4.89	0.7782
	137	116	253	
Total	5.03	5.10	5.06	0.5618
	318	241	559	
t-test (p)	0.1009	0.0775	0.0177	

B's expectations of trust on a 7-point Likert scale

This table also shows that while the differences across years are insignificant in any treatment at conventional levels, there is overall a significant reduction in expectations of trust overall. These data suggest that while the treatment effect in Y1 is probably not due to guilt aversion, the reduction in transfers between Y1 and Y3 may have be related to reduced guilt. This is not particularly at odds with the spirit of the argument in the paper, but it should be accounted for. Indeed, in table 4 it was found to be a highly significant predictor of the percent transferred, as guilt aversion would imply.

To further test this, we re-ran Regression 1 from Table 4 without the expectations control, and performed a seemingly unrelated regressions (SUR) test on the change in the coefficients. Interestingly, removing the expectations as a covariate generates a gender effect. However, it has no substantial effect on the main treatment variables either in magnitude or significance. The SUR test of the effect on the first three covariates has a Chi-square(3) p-value of 0.8288, indicating that they are not significantly different in the two specifications. Therefore, to the extent that expectations of trust are a good proxy for second-order beliefs, the mediation of the treatment effect by any "moral disengagement" over the period seems insignificant.

IV. A MODEL

The results above indicate that Y3 students feel much less bad about their lies than do Y1. However, this explanation does not seem to fit with the entirety of the data, as it would imply that they should lie more, whereas empirically they lie less. Moreover, the reduction in lies is itself predicted to reduce the degree of dissonance, and so may be part of the explanation for the difference in transfer rates. This raises the intriguing possibility that Y3 students may be the same as Y1 in their basic preferences, but more “foresighted”. They reduce their lies in the first stage of the experiment enough to justify lower transfers later on. In essence, they become *sophisticated liars*. In the terms of moral balancing, they engage in less moral cleansing in their transfer behavior because they manage to do some moral licensing in the lie phase. This is a more complex phenomenon, and in this section we develop a simple theoretical model to explain the pattern of behavior. We emphasize that this modeling work is *post hoc* explanation, rather than predictive. It is interesting because it highlights another facet of how participation in the program can at the same time affect behavior, and leave basic preferences untouched. That is, the model of increasing sophistication represents another angle on the mixed results of selection and indoctrination effects outlined in the introduction. At an informal level, it is also very much in line with the result that the Y3 students managed to earn more money from the Roll condition than did the Y1. By the nature of optimization, after all, one would expect increasing sophistication to allow decision makers to economize on their moral balancing.

The model has three basic sets of assumptions. First, there is a mechanic of *moral credit*, a utility argument that reflects self-image benefit, which people trade off against material benefit. Second, there is a distinction between two *types* of individual, specifically *sophisticated types*, who maximize total utility, and *naïve* types who only maximize over one of two variables in the first stage. Finally, we assume the existence of a set of *personal norms* that determine the maximum declared roll or keep share that does not generate moral harm. The main goal of the model is to explain the difference across years in the Roll condition, showing how “sophistication” may have shifted both transfer rates and lie rates in different directions. It also illustrates the importance of social norms in the reduction of moral balancing.

We begin with a population of decision makers, acting over time. Each is endowed at any moment *t* with a certain, potentially idiosyncratic *moral credit* K_t, and must take actions a_t that have utility benefit *u* and moral cost *v*. The source of the cost is essentially to run down the moral credit. To motivate this intuition, we note that a more detailed model, along the lines of Bénabou and Tirole (2016) might treat this cost as a *self-signalling* effect. According to this argument, people get some self-image benefit from their beliefs about their own moral type, but do not have perfect information about what that type is. Instead, their self-image is based on the Bayesian posterior expectation of that type, given previous behavior. In this extended model, *K* would be the expected value of one's own moral type, from which one derives a self-image benefit, and the cost would reflect the updating to this belief based on the equilibrium choices of *a*. That is, utility at time *t* would be defined as

$$U_t = u(a_t) + \theta\big[E[K \vee K_{t-1}, a_t]\big]$$

Where the increasing function θ is the utility effect of the self-image. However, for reasons of tractability we will not explicitly model the beliefs. Rather, we simplify by writing the image effect of the action as a cost corresponding to the shift in posterior beliefs that it engenders:

$$U_t = u(a_t) + V[K_{t-1} - v_a(a_t)]$$

The *signal cost*, *v*, is an informational effect. The moral *self-image benefit* is encapsulated in *V*, and reflects how the decision maker feels about this shift. We assume $v' > 0$, $v'' > 0$, $V' > 0$ and $V'' < 0$. So increasing the action has increasing marginal signalling cost, and the utility benefit of a good moral self-image has decreasing marginal benefit.

The optimal action will satisfy the first-order condition

$\frac{\partial U}{\partial a_t} = u(a_t) - V[K_{t-1} - v_a(a_t)]v(a_t) = 0$

The marginal benefit of the action is set equal to its marginal moral cost, where this latter is defined as the product of the marginal signal cost and the marginal self-image benefit that this signal diminishes. It is immediate from this set-up that anything that reduces the base level of moral credit K_{t-1} will thereby increase the self-image cost of the action, and thus reduce the level chosen. This includes, for instance, a_{t-1}, and is the moral balancing result. *A morally costly action in period t will reduce the level at which subsequent actions are taken at* $t+1$.

Made more specific for this experiment, we assume that the decision maker comes to the problem with some pre-existing level of moral self-worth K_0, and rewrite the model

$$U = 4(r + (6 - r)l)\frac{(k+1)}{2} + V[K_0 - v_l(l) - v_k(k)] \qquad (1)$$

This formulation may require some explanation, as it introduces the second element of the model, the *personal norm*. The parameter r is the number rolled, and l and k, which respectively indicate *lying* and *keeping* and are both bounded between 0 and 1, are the *normatively relevant* choices.[8] The form of the first term after the equals sign defines a lie as a declaration between r and 6. Implicitly, this assumes that there is no moral cost for any declaration less than or equal to the roll. Similarly, the form of k implies that there is no moral cost to transferring more than half of the endowment. Underreporting the roll and offering "supergenerous" transfers are therefore dominated actions, which the form of (1) rules out by assumption.

The idea that there is no moral cost to an honest report is probably uncontroversial, as is the presumption that the great majority of deviations from this "honesty norm" are positive. The implicit norm of an "equal split" of the pie is somewhat less clear. Empirically, however, we saw

[8] For clarity of exposition, we model these as continuous.

in Figure 3 above that the fraction of participants who did offer more than half of the endowment is vanishingly small. Less than 2% (11 out of 578) did so, in fact, more or less equally spread across the Year-Roll conditions, a number we attribute to noise. The evidence of a norm that allows transfers to be more than half of the pie seems overwhelming.

Honesty and equal splits are only special cases of the norms that could exist. More generally, assume there is a norm n_d for declarations d, and one n_s for transfers s, and that both represent thresholds for acceptable behavior. Then expression (1) becomes

$$U = 4(n_d + (6 - n_d)l)(n_t + (1 - n_t)k) + V[M_0 - v_l(d) - v_k(1 - s)]$$

With $l = (d - n_d)/(6 - n_d)$ and $k = (1 - s - n_t)/(1 - n_t)$ as the extent of the normatively relevant actions (above the thresholds). The signal costs depend on the Bayesian interpretation of the actual declaration and transfer, and should therefore have general slopes independent of the personal norm, but their importance depends on the norm. We make the following assumption

Free action below the norm: For each action $a = l, k$, $v_a(a) = 0$ if $a < 0$

That is, there is no moral cost for actions taken below the threshold, but the signal value of the action picks back up for actions above. The key element of this assumption is that the marginal cost of actions above the threshold is therefore independent of the threshold itself. If norms differ between individuals, those with high and low norms will still agree on the signal cost of a particular action. However, each can engage in the action up to her own personal norm at no moral cost. Given a set of norms, the level makes little qualitative difference to the results below.

However, it does make a quantitative difference. For most of the development below, we will focus on the case in expression (1). The honesty norm is so natural that although the double-blind experiment does not allow us to check empirically, if the reader is like us, she didn't even think of it as a norm at first. The even-split norm's empirical relevance has already been mentioned.

We now introduce the third element of the model, the concept of *sophisticated* and *naïve* types. Type is independent of K_0 for simplicity; types differ only in the process by which they solve the problem. The intuition is one of discount rates. The choice *l* is made in phase 1 of our experiment, and the choice *k* in phase 2. Naïve types have a "discount" of 100%, and do not consider the effect of period-2 action on their moral self-image when they make the first choice. Sophisticated types have a discount rate of 0, and therefore maximize utility from expression (1) for both periods together. We economize on notation, and model this by dropping the image cost of keeping money for the period-1 decision for naïve types. To continue the analogy to the basic consumer problem, imagine that the goods are purchased sequentially. Naïve types do not consider the optimization over both goods when they buy the first. Standing in the bakery, say, they "forget" that they need to go to the grocer's later on, and spend all their "funds" on cake.

The first-order conditions for the sophisticated types are

$$4(6-r)\,(k^s+1)/2 = V[K_0 - v_l(l^s) - v_k(k^s)]v_l(l^s) \tag{2}$$

$$4\,(r+(6-r)l^s)/2 = V[K_0 - v_l(l^s) - v_k(k^s)]v_k(k^s) \tag{3}$$

These yield the familiar condition that the ratio of marginal benefits to the two actions should be the inverse of the ratio of marginal costs, implicitly defining the optimal level of each.

$$\frac{(6-r)(k^s+1)}{r+(6-r)l^s} = \frac{v_l(l^s)}{v_k(k^s)} \tag{4}$$

Several points about expression (4) deserve attention. First, while it appears that for sophisticated types, the optimal ratio of *l* to *k* is independent of the pre-existing level of moral credit K_0, it should be kept in mind that the ratio on the right-hand side is nonlinear marginal signalling costs, and so in principle could vary as the overall "moral budget" is relaxed. On the other hand, (4) can easily be shown to represent a monotonic increasing relationship between *k* and *l*; sophisticated types with greater moral credit in this model will "spend" it on lying and keeping.[9] Second, the left-hand side falls in *r*, the actual roll. The intuition is that *r* increases the size of the pie, and therefore also the marginal benefit of *k*, but reduces the possible scope of the lie, and with it the marginal benefit of *l*, since a given percentage increase in a smaller scope of potential lies represents less money. As a result, the equalization of the left- and right-hand ratios implies less *l* "per unit of *k*" as *r* rises. Those who roll higher, lie proportionately less. Third, we note that ceteris paribus, the optimal mix of *k* and *l* depends inversely on the message costs of the two actions. This fact represents a "demand curve" for morally costly behavior; the higher the relative cost of either activity, the lower the degree engaged in. Finally, the monotonic relationship between *k* and *l* at the optimum implies a unique solution, found by re-introducing the "budget constraint," that is, maximizing (1) with respect to one variable, subject to the other satisfying the relationship in (4).

Naïve types, by contrast, make lie decisions without respect for later keeping. In phase one, they treat *k* as a default value, denoted $\acute{k}$. Because it is not a choice they consider in the first decision,

[9] This may seem like a counterintuitive result. Those who see themselves as "better people" end up acting worse! The model as written abstracts from – or more charitably holds constant – an arguably important aspect of moral behavior that would attenuate this effect, which is the level of moral credit to which one aspires. Given that everyone by assumption (specifically on the constant form of *V* across individuals) wants to be the same; those above the target therefore have more "moral latitude".

naïve types exclude the signal cost of keeping from the function *V*. The first-order conditions for the naïve types are

$$4(6-r)\left(\acute{k}+1\right)/2 = V[K_0 - v_l(l^n)]v_l(l^n) \quad (5)$$

$$4(r+(6-r)l^n) = V[K_0 - v_l(l^n) - v_k(k^n)]v_k(k^n) \qquad (6)$$

Comparing the forms of (2) and (5) shows the effect of naiveté on lying. Consider expression (7) below. The first term to the right of the equal sign is the condition on (naively) optimal marginal cost of lies. The second term on the right is the condition that would hold for equal *k* if the decision maker were sophisticated.

$$v(l^n) = \frac{4(6-r)\left(\acute{k}+1\right)}{V[M_0 - v_l(l^n)]} > \frac{4(6-r)\left(\acute{k}+1\right)}{V\left[M_0 - v_l(l^n) - v_k\left(\acute{k}\right)\right]} \qquad (7)$$

Ignoring the future moral cost of keeping, as per expression (5), results in basing the lie on a larger moral credit than will turn out to be the case, as can be seen in the denominators in (7). The form of *V* therefore implies a lower marginal cost of self-image loss for naïve types, which can be compensated with a higher marginal image cost of lying than would be chosen by sophisticated types with equal *k*. That is, *naïve types lie more than sophisticated for a given expectation of k.* This is quite intuitive; the naïve types are “tempted” by the gain from lying, and by construction do not foresee that they will have to “pay for” their “transgression” with lower transfers later, so they lie “too much” from the perspective of sophisticated types.[10] In short, even if $\acute{k}$ is based on expectations of the sophisticated amount of keeping k^s, naïve types will still engage in “too much lying” in phase 1, and have to compensate with “too little keeping” in phase 2. To show that this payment actually occurs in the model, note that expressions (3) and (6) have the same form; only the values of l^n versus l^s drive the difference in the percentage kept. Having determined *l*, this

[10] Of course, $\acute{k}$ is a free parameter not determined within the model. Our basic assumption is that $k^s \geq \acute{k}$.

expression identifies the corresponding *k* on the “moral budget line”. Note that this is not entirely unambiguous: while it is true that a higher phase-1 lie rate reduces the phase-2 moral credit (making keeping more costly) it also increases the endowment, making keep rates more lucrative. To more rigorously identify the sign of the relationship, we therefore make use of the implicit function *k*(*l*). Suppressing superscripts since the relationship applies to both types, this gives

$$4\,(r+(6-r)l)/2 = V\big[K_0 - v_l(l) - v_k\big(k(l)\big)\big]v_k\big(k(l)\big)$$

Further suppressing the arguments for expositional clarity and taking the differential with respect to *l* then yields

$$2(6-r) = v_k\frac{dV'}{dl} + V'\frac{dv_k{}'}{dl}$$

$$2(6-r) = v_k\left[-V''\left(v_l' + v_k{}'\frac{dk}{dl}\right)\right] + V'v_k{}''\frac{dk}{dl}$$

Finally, isolating *dk*/*dl*,

$$\frac{dk}{dl} = \frac{2(6-r)+V''v_l'v_k'}{V'v_k''-V''v_k'v_k'} \qquad (8)$$

The denominator of (8) is unambiguously positive. Indeed, it is negative value of the second-order condition of *U* with respect to *k*. The numerator is the cross-partial derivative of *U*. We show in the appendix that for lies above the optimal amount from a sophisticated perspective, *dk/dl* must be negative. Thus, having lied more, naïve types later rationally decide to keep less, consistent with our moral balancing argument. We also show that whether the absolute slope of the relationship is greater or less than unity depends on the difference between the signal cost of lies, v_l and that of keeping, v_k. Because the marginal benefit of *k* is increased relative to *l* by the value of the honest roll *r*, the condition $v_l > v_k$ is necessary for the slope to be greater than 1 in absolute value. The main theoretical result is therefore

Proposition: Naïve types lie more and keep less than sophisticated types. The relative shifts depend on the difference between the costs of each action.

How does this model compare with the data from the experiment? The key values are as shown in Table 9, using the estimated average lies calculated above, and the "normatively relevant" keep rates defined by the percent of the pie kept above 0.5

	l	*k*
Y1	0.732	0.238
Y3	0.576	0.462

Average normatively relevant keep and lie rates from the experimental results

Replacing the value of *r* with its average of 3.5 and substituting into the equation (4) suggests that at the Y3 "optimum" point, the ratio of the marginal signal cost of lying to the marginal signal cost of keeping is nearly unity.

$$\frac{v_l(l^s)}{v_k(k^s)} = \frac{(6-r)(k^s+1)}{r+(6-r)l^s} = \frac{(2.5)(0.462+1)}{3.5+(2.5)0.555} = 1.04$$

Therefore, on the margin, the empirical estimate of the signal costs is about the same for each action, which is what one would expect at an optimum. If we estimate the slope of the "moral budget curve" using the two points available, we find

$$\frac{dk}{dl} = \frac{0.462 - 0.238}{0.576 - 0.732} = -1.436$$

As mentioned above, the condition that this slope is greater than unity implies that over the range between the points, $v_l' > v_k'$. This is consistent with the model's suggestion that between the Y3 and Y1 chosen points, where lying was higher and keeping was lower, the marginal cost of the former was greater than that of the latter, reflecting the sub-optimality of naïve strategies.

As a final point of contact between the model and the empirical results, note that within the model as written, naïve types have downward pressure on their *keep* rates due to excessive lying in the first phase, while sophisticated types do not. It is possible that this depression of keeping on naïve types could push them into a corner solution, which in our model is at *k* = 0.5. This is harder to explain among sophisticated types. Therefore, a prediction of the model is that there should be more even-split offers among Y1 participants than among Y3 in the *Roll* condition, but not in the *No-Roll*. The table below shows that this is indeed the case. In the *Roll* condition, 49.6% of Y1 participants transferred half of their endowment; among the Y3 participants, the rate was less than half of that, at 21.8%. This difference is significant (Chi-square (1) p = 0.000). In the *No-Roll*, by contrast, the difference was insignificant: 30.6% in Y1 versus 26.6% in Y3 (Chi-square (1) p = 0.433).

TABLE 10

year	***No-Roll***	***Roll***	**Total**
Y1	0.306	0.496	0.381
N	*193*	*127*	*320*
Y3	0.266	0.218	0.244
N	*139*	*119*	*258*
Total	0.289	0.362	0.320
N	*332*	*246*	*578*

Proportion of each year and treatment condition transferring exactly half of the endowment. N = number of observations.

This model of increasing sophistication therefore seems to conform to the empirical patterns across years in the *Roll* condition quite well, although of course other explanations are possible for any of the particular empirical results discussed. It also matches the observed data across treatments in Y1. In the *No-Roll* treatment, the "declaration" variable should impose no moral cost at all, as it does not require lies. Therefore, the model suggests *Keep* rates should be higher – that is, transfers lower – in the *No-Roll* than in the *Roll* condition, and Y1 transfers do indeed diminish across the treatment variable. However, it should also predict a difference across treatments in Y3, whereas in the data the transfers were identical for that year. In effect, the model does not predict the reduction – evaporation – of moral balancing that we see empirically. Here we come back to the question of norms. Consider what happens if the norm happens to be equal to the sophisticated maximizing behavior. Now the maximizing behavior even in the *Roll* condition imposes no moral cost by assumption, but the cost functions pick back up in the case of exceeding it, keeping sophisticated behavior stable. Therefore, the *No-Roll* condition results in identical behavior, for the same distribution of multipliers.[11]

IV. DISCUSSION

Taking stock of our results, we have found a robust effect by which Y1 participants in the B-role lie to a greater extent than do those in Y3. When they do, but only when they do, they also

[11] This informal discussion leaves several features unaccounted for. For instance, there might be an iterative process of normalization and re-maximization. We conjecture that depending on the mathematical structure of the costs and beliefs, this could either look like a "contraction" and result in an interior solution, or may very well result in a full unravelling of the social norm, and a corner solution.

transfer back more to A. In the absence of dishonesty, and controlling for expectations and endowment, the transfers are not different between the years, although there seems to be some difference in the social norms as measured through our proxy for second-order beliefs.

This appears to be a rather grim reckoning. Like several earlier studies, we find evidence of selection due to the high level of lying among Y1 participants and the constant distributional preferences revealed in the *No-Roll* condition. And with regard to the changes that do occur over time, exactly what do our participants seem to be learning? On the one hand, participants lie somewhat less in third year, although lie rates are still very high compared to similar studies. On the other hand, if anything, the level of reciprocity falls, which appears to be a movement towards selfishness. And the moral weight of the lie – the change in the difference between treatment conditions across years – seems to vanish altogether. If we then consider that the net effect is an increase in monetary payoff over time in the *Roll* condition that does not appear in the *No-Roll*, it appears that even the restriction of lying itself may have self-interested purposes.

To think through the last point more clearly, we developed a model of increasing "moral sophistication" in this sense. It yielded predictions that fit the change across years – and across treatments in Y1 – relatively well. The data from this experiment are consistent with a story in which, by their third year in the business school, students have become more sophisticated liars, restricting their dishonesty in the first stage of their decision enough to then not feel so much dissonance that they have to pay for it in the second stage.

Or almost consistent. This model of increasing sophistication assumes that the moral pressure remains active. It seems inconsistent to say on the one hand (in the Roll condition, comparing Y3 and Y1) that Y3 are optimizing their moral balancing better, and on the other hand (in Y3, comparing Roll and No-Roll) that there is no evidence of moral balancing at all! We finished the theoretical section with a suggestion in this regard of changes in personal norms. While the discussion was informal, the idea is that behavior that is “good enough” imposes no moral cost on the decision maker. What Y3 students have learned, in this story, is that maximizing behavior constitutes that “good enough” behavior. Thus the personal norm has come into alignment with the originally transgressive social norms of optimization, and no longer imposes any cost.

This phenomenon is more complex than the change in moral balancing over time that constituted our original hypotheses, and such a model is incomplete without a mechanism that describes how the norms are established and change. “Explaining” differences in behavior by invoking different norms begs the question to some extent. Future work must crucially identify a mechanism for norm internalization to make the story sound. However, this seems like a promising lead. The story fits the data, and links the selection and indoctrination literatures in a natural way. Consistent with the selection literature, studying in a business school does not change basic preferences; students learn to be more foresighted in their actions, and adopt the maximizing behavior as normatively “justified”. In this way, it corresponds to our original goal of describing *the socialization of maximizing behavior*.

We close by noting as a limitation of the study that our assumption has been that differences in the transfer rates across the *Roll* variable have been assumed to measure moral balancing, but that links other than this may well exist. For instance, it is conceivable that participants who “worked for” the money by rolling a die would feel more entitled to keep it. This, however, has the opposite prediction, which is lower transfers in *Roll* than in *No-Roll*, so no confound is possible there. As a second concern, the value of M is endogenous in the *Roll* condition but not in the *No-Roll*, which implies potential selection effects. If lying depends on some heterogeneous, unobservable type, then it would be safe to assume that the people who get $M = 6$ in the *Roll* condition are different along that dimension than those who get the same value of *M* in the *No-Roll* condition. The prediction then turns on the correlation between this unobservable type and sharing behavior. If those who have a greater tendency to lie also have a greater propensity for sharing, then it could be this factor, rather than the moral balancing, that explains the pattern of greater sharing in the *Roll* than *No-Roll* condition. We would argue that lying costs are more likely to be positively than negatively correlated with pro-social distributional concerns, with some "prosocial" types who both share and report truthfully, and other "selfish" types who do neither. This would predict that those who lie share *less*, not more, than those who report honestly, and would also push in the opposite direction than the moral balancing argument. But we cannot necessarily reject the possibility that another effect is interfering here. We also note that a more direct test of the hypothesis of increasing sophistication might compare the "partial strategy" method employed here with a "full strategy" method, in which B must declare a transfer for every possible value of the multiplier before rolling (or seeing) it. This would push B to consider both factors at once, and so should bring results back into line with the sophisticated strategy. These issues, as well as direct comparison with a control group from a different discipline, represent useful avenues for future research.

BIBLIOGRAPHY

AACSB. 2004. *Ethics Education in Business Schools: Report to the Ethics Education Task Force*. AACSB International – The Association to Advance Collegiate Schools of Business. 22 p.

Abeler, Johannes, Daniele Nosenzo, and Collin Raymond. 2016. “Preferences for Truth-Telling.”

Abend, Gabriel. 2013. “The Origins of Business Ethics in American Universities, 1902–1936.” *Business Ethics Quarterly* 23 (2): 171–205.

Akrivou, Kleio, and Hilary Bradbury-Huang. 2015. “Educating Integrated Catalysts: Transforming Business Schools toward Ethics and Sustainability.” *Academy of Management Learning & Education* 14 (2): 222–240.

Andel, Chantal EE van, Joshua M. Tybur, and Paul AM Van Lange. 2016. “Donor Registration, College Major, and Prosociality: Differences among Students of Economics, Medicine and Psychology.” *Personality and Individual Differences* 94: 277–283.

Barr, Abigail, and Georgia Michailidou. 2017. “Complicity without Connection or Communication.” *Journal of Economic Behavior & Organization* 142: 1–10.

Bauman, Yoram, and Elaina Rose. 2011. “Selection or Indoctrination: Why Do Economics Students Donate Less than the Rest?” *Journal of Economic Behavior & Organization* 79 (3): 318–327.

Bekkers, René, and Pamala Wiepking. 2011. “A Literature Review of Empirical Studies of Philanthropy: Eight Mechanisms That Drive Charitable Giving.” *Nonprofit and Voluntary Sector Quarterly* 40 (5): 924–973.

Berg, Joyce, John Dickhaut, and Kevin McCabe. 1995. “Trust, Reciprocity, and Social History.” *Games and Economic Behavior* 10 (1): 122–142.

Bowles, Samuel. 2016. *The Moral Economy: Why Good Incentives Are No Substitute for Good Citizens*. Yale University Press.

Burns, David J., James A. Tackett, and Fran Wolf. 2015. “The Effectiveness of Instruction in Accounting Ethics Education: Another Look.” In *Research on Professional Responsibility and Ethics in Accounting*, 149–180. Emerald Group Publishing Limited.

Carter, John R., and Michael D. Irons. 1991. “Are Economists Different, and If so, Why?” *Journal of Economic Perspectives* 5 (2): 171–177.

Caviola, Lucius, and Nadira Faulmüller. 2014. “Moral Hypocrisy in Economic Games—how Prosocial Behavior Is Shaped by Social Expectations.” *Frontiers in Psychology* 5: 897.

Cipriani, Giam Pietro, Diego Lubian, and Angelo Zago. 2009. “Natural Born Economists?” *Journal of Economic Psychology* 30 (3): 455–468.

Clot, Sophie, Gilles Grolleau, and Lisette Ibanez. 2013. “Moral Self-Licensing and Social Dilemmas: An Experimental Evidence from a Taking Game in Madagascar.”

Delis, Manthos D., Iftekhar Hasan, and Maria Iosifidi. 2017. “On the Effect of Business and Economic University Education on Political Ideology: An Empirical Note.” *Journal of Business Ethics*, 1–14.

Dhami, Sanjit. 2018. “Human Ethics and Virtues: Rethinking the Homo-Economicus Model.”

Etzioni, Amitai. 2015. “The Moral Effects of Economic Teaching.” In *Sociological Forum*, 30:228–233. Wiley Online Library.

Festinger, L. 1957. “1957 A Theory of Cognitive Dissonance. Stanford, Calif.: Stanford

University Press."

Fischbacher, Urs, and Franziska Föllmi-Heusi. 2013. "Lies in Disguise—an Experimental Study on Cheating." *Journal of the European Economic Association* 11 (3): 525–547.

Frank, Björn, and Günther G. Schulze. 2000. "Does Economics Make Citizens Corrupt?" *Journal of Economic Behavior & Organization* 43 (1): 101–113.

Frank, Robert H., Thomas Gilovich, and Dennis T. Regan. 1998. "Does Studying Economics Inhibit Cooperation." *Economics, Ethics, and Public Policy* 1 (2): 51.

Frey, Bruno S., and Stephan Meier. 2003. "Are Political Economists Selfish and Indoctrinated? Evidence from a Natural Experiment." *Economic Inquiry* 41 (3): 448–462.

———. 2005. "Selfish and Indoctrinated Economists?" *European Journal of Law and Economics* 19 (2): 165–171.

Frey, Bruno S., Werner W. Pommerehne, and Beat Gygi. 1993. "Economics Indoctrination or Selection? Some Empirical Results." *The Journal of Economic Education* 24 (3): 271–281.

Gerlach, Philipp. 2017. "The Games Economists Play: Why Economics Students Behave More Selfishly than Other Students." *PloS One* 12 (9): e0183814.

Gerlach, Philipp, Kinneret Teodorescu, and Ralph Hertwig. 2017. "The Truth about Lies. A Meta-Analysis on Dishonest Behavior." *Manuscript in Preparation*.

Ghoshal, Sumantra. 2005. "Bad Management Theories Are Destroying Good Management Practices." *Academy of Management Learning & Education* 4 (1): 75–91.

Giacolone, R. A. (2015) *Business Ethics and Management Education*. Virtual collection in Academy of Management Publications.

Grossman, Zachary. 2015. "Self-Signaling and Social-Signaling in Giving." *Journal of Economic Behavior & Organization* 117: 26–39.

Haucap, Justus, and Andrea Müller. 2014. "Why Are Economists so Different? Nature, Nurture and Gender Effects in a Simple Trust Game."

Houser, D., Schunk, D. (2009). "Social environments with competitive pressure: Gender effects in the decisions of german schoolchildren*." Journal of Economic Psychology* 30 (4): 634–641.

Hummel, Katrin, Dieter Pfaff, and Katja Rost. 2016. "Does Economics and Business Education Wash Away Moral Judgment Competence?" *Journal of Business Ethics*, 1–19.

Irlenbusch, Bernd, and Marie Claire Villeval. 2015. "Behavioral Ethics: How Psychology Influenced Economics and How Economics Might Inform Psychology?" *Current Opinion in Psychology* 6: 87–92.

Jacobsen, Catrine, Toke Reinholt Fosgaard, and David Pascual-Ezama. 2018. "Why Do We Lie? A Practical Guide to the Dishonesty Literature." *Journal of Economic Surveys* 32 (2): 357–387.

Johnson, Noel D., and Alexandra A. Mislin. 2011. "Trust Games: A Meta-Analysis." *Journal of Economic Psychology* 32 (5): 865–889.

Joule, Robert-Vincent, and Jean-Léon Beauvois. 1997. "Cognitive Dissonance Theory: A Radical View." *European Review of Social Psychology* 8 (1): 1–32.

Kahneman, Daniel, Jack L. Knetsch, and Richard Thaler. 1986. "Fairness as a Constraint on Profit Seeking: Entitlements in the Market." *The American Economic Review*, 728–741.

Kirchgässner, Gebhard. 2005. "(Why) Are Economists Different?" *European Journal of Political Economy* 21 (3): 543–562.

Konow, James. 2014. "Can Economic Ethics Be Taught." Discussion Paper, University of Kiel.

Krick, Annika, Stephanie Tresp, Mirijam Vatter, Antonia Ludwig, and Michael Wihlenda. 2016. "The Relationships Between the Dark Triad, the Moral Judgment Level, and the Students' Disciplinary Choice." *Journal of Individual Differences*.

Laband, David N., and Richard O. Beil. 1999. "Are Economists More Selfish than Other'social'scientists?" *Public Choice* 100 (1–2): 85–101.

Lewis, Alan, Alexander Bardis, Chloe Flint, Claire Mason, Natalya Smith, Charlotte Tickle, and Jennifer Zinser. 2012. "Drawing the Line Somewhere: An Experimental Study of Moral Compromise." *Journal of Economic Psychology* 33 (4): 718–725.

López-Pérez, Raúl, and Eli Spiegelman. 2012. "(Why) do economists lie more?" Forthcoming in *Deception in Behavioral Economics*, Elsevier.

Marwell, Gerald, and Ruth E. Ames. 1981. "Economists Free Ride, Does Anyone Else?: Experiments on the Provision of Public Goods, IV." *Journal of Public Economics* 15 (3): 295–310.

Meier, Stephen, and Bruno S. Frey. 2004. "Do Business Students Make Good Citizens?" *International Journal of the Economics of Business* 11 (2): 141–163.

Merritt, Anna C., Daniel A. Effron, and Benoît Monin. 2010. "Moral Self-Licensing: When Being Good Frees Us to Be Bad." *Social and Personality Psychology Compass* 4 (5): 344–357.

Meub, Lukas, Till Proeger, Tim Schneider, and Kilian Bizer. 2016. "The Victim Matters–Experimental Evidence on Lying, Moral Costs and Moral Cleansing." *Applied Economics Letters* 23 (16): 1162–1167.

Mullen, Elizabeth, and Benoît Monin. 2016. "Consistency versus Licensing Effects of Past Moral Behavior." *Annual Review of Psychology* 67.

Ostrom, E., 2000. Collective action and the evolution of social norms. *Journal of Economic Perspectives*, *14*(3), pp.137-158.

Ploner, Matteo, and Tobias Regner. 2013. "Self-Image and Moral Balancing: An Experimental Analysis." *Journal of Economic Behavior & Organization* 93: 374–383.

Racko, Girts, Karoline Strauss, and Brendan Burchell. 2017. "Economics Education and Value Change: The Role of Program-Normative Homogeneity and Peer Influence." *Academy of Management Learning & Education* 16 (3): 373–392.

Rosenbaum, Stephen Mark, Stephan Billinger, and Nils Stieglitz. 2014. "Let's Be Honest: A Review of Experimental Evidence of Honesty and Truth-Telling." *Journal of Economic Psychology* 45: 181–196.

Rubinstein, Ariel. 2006. "A Sceptic's Comment on the Study of Economics." *The Economic Journal* 116 (510).

Sachdeva, Sonya, Rumen Iliev, and Douglas L. Medin. 2009. "Sinning Saints and Saintly Sinners: The Paradox of Moral Self-Regulation." *Psychological Science* 20 (4): 523–528.

Seçilmiş, Erdem. 2018. "An Experimental Analysis of Moral Self-Regulation." *Applied Economics Letters* 25 (12): 857–861.

Shalvi, Shaul, Jason Dana, Michel JJ Handgraaf, and Carsten KW De Dreu. 2011. "Justified Ethicality: Observing Desired Counterfactuals Modifies Ethical Perceptions and Behavior." *Organizational Behavior and Human Decision Processes* 115 (2): 181–190.

Shalvi, Shaul, and David Leiser. 2013. "Moral Firmness." *Journal of Economic Behavior & Organization* 93: 400–407.

Sigurjonsson, Throstur Olaf, Audur Arna Arnardottir, Vlad Vaiman, and Pall Rikhardsson. 2015. "Managers' Views on Ethics Education in Business Schools: An Empirical Study."

Journal of Business Ethics 130 (1): 1–13.

Smith, Adam. 2010. [1759] *The Theory of Moral Sentiments*. Penguin.

Smith, V.L., 2003. Constructivist and ecological rationality in economics. *American economic review*, *93*(3), pp.465-508.

Wang, Long, Deepak Malhotra, and J. Keith Murnighan. 2011. “Economics Education and Greed.” *Academy of Management Learning & Education* 10 (4): 643–660.

Wang, Long, Chen-Bo Zhong, and J. Keith Murnighan. 2014. “The Social and Ethical Consequences of a Calculative Mindset.” *Organizational Behavior and Human Decision Processes* 125 (1): 39–49.

Whitehead, Alfred North. 2011. *Science and the Modern World*. Cambridge University Press.

Yezer, Anthony M., Robert S. Goldfarb, and Paul J. Poppen. 1996. “Does Studying Economics Discourage Cooperation? Watch What We Do, Not What We Say or How We Play.” *Journal of Economic Perspectives* 10 (1): 177–186.

Appendix :

Show: Sophisticated types who lie more also keep more.

At the optimum, the sign of relationship between *k* and *l* can be found by differentiating the implicit relationship in expression (4) in the text with respect to one variable.

From (4)

$$\frac{(6-r)k^s}{r+(6-r)l^s} = \frac{v_l(l^s)}{v_k(k^s)}$$

Define A = (6 – *r*), and drop arguments when possible

$$\frac{Ak^s}{r+Al^s} = \frac{v_l(l^s)}{v_k(k^s)}$$

$$k^s = \frac{(r+Al^s)}{A}\frac{v_l(l^s)}{v_k(k^s)}$$

$$1 = \frac{dl}{dk}\left(\frac{v_l(l^s)}{v_k(k^s)} + \frac{(r+Al^s)}{A}\frac{v_l''v_k'}{(v_k)^2}\right) - \frac{(r+Al^s)}{A}\frac{v_l'v_k''}{(v_k)^2}$$

$$\frac{dl}{dk} = \left(1 + \frac{(r+Al^s)}{A}\frac{v_l'v_k''}{(v_k)^2}\right)\left(\frac{v_l(l^s)}{v_k(k^s)} + \frac{(r+Al^s)}{A}\frac{v_l''v_k'}{(v_k)^2}\right)^{-1} > 0$$

Effect of increased lies on keeping

Show: In period 2, a decision maker with equal r and M_0 will keep less, if the lie was bigger.

First fix *l* and find the function that determines keeping. Writing the first order condition that determines keeping as an implicit function of lies

$$4(r+(6-r)l^s)-V\left[M_0-v_l(l^s)-v_k\left(k^s(l)\right)\right]v_k\left(k^s(l)\right)=0 \qquad \text{(A1)}$$

Now see how this changes with *l*

Differentiating with respect to *l* and solving for dk/dl gives

$$4(6-r)-\left(V''\left(-v_l'-v_k{}'\frac{dk}{dl}\right)v_k+Vv_k{}''\frac{dk}{dl}\right)=0$$

$$4(6-r)+\left(V''\left(v_l'+v_k{}'\frac{dk}{dl}\right)v_k-Vv_k{}''\frac{dk}{dl}\right)=0$$

$$4(6-r)+\left(V''v_l'v_k+V''v_k{}'\frac{dk}{dl}v_k-Vv_k{}''\frac{dk}{dl}\right)=0$$

$$4(6-r)+V''v_l'v_k+(V''v_k{}'v_k-Vv_k{}'')\frac{dk}{dl}=0$$

$$4(6-r)+V''v_l'v_k=(Vv_k{}''-V''v_k{}'v_k{}')\frac{dk}{dl}$$

$$\frac{4(6-r)+V''v_l'v_k'}{V'v_k''-V''\left(v_k'\right)^2}=\frac{dk}{dl}=\frac{-\partial^2U/\partial l\partial k}{\partial^2U/\partial k^2} \qquad \text{(A2)}$$

The denominator is positive, the negative of the direct second-order condition on lies. The numerator is the cross partial of the utility function. The problem is that *k* and *l* are complements in benefit, but substitutes in cost. We go back to the first-order condition on keeping,

$$4(r+(6-r)l^s)-V[M_0-v_l(l^s)-v_k(k^s)]v_k(k^s)=0$$

And define *Q* for compactness

$$4(r+(6-r)l^s)-Q(l^s,k^s)=0 \qquad \text{(A3)}$$

Because this represents marginal utility at the optimum, increasing either *k* or *l* will reduce the value. A first-order Taylor expansion just above the optimum therefore gives

$$4\left(r+(6-r)l^s+(6-r)(x-l^s)\right)-(Q(l^s,k^s)+(x-l^s)Q_l+(y-k^s)Q_k)<0$$

Which implies with (A3)

$$4(6-r)(x-l^s)-(x-l^s)Q_l-(y-k^s)Q_k<0 \qquad \text{(A4)}$$

Now

$$Q = V[M_0 - v_l(l^s) - v_k(k^s)]v_k'(k^s)$$

$$Q_l = -V''v_k'v_l' > 0 \tag{A5}$$

$$Q_k = -V''{v_k'}^2 + V'v_k'' > 0 \tag{A6}$$

If we fix *k* and look at the change only in *l*, then $y = k^s$, and by subsutituting (A5) and (A6) into (A4), we get

$$4(6-r) - (-V''v_k'v_l') < 0$$

$$4(6-r) + V''v_k'v_l' < 0 \tag{A7}$$

Which gives the sign on the numerator of (A2)

Difference in costs determines the slope of the effect –

Show: the response favors the cheap action

Begin with the response of *k* to a change in *l*, from (A2). We want to see the condition such that

$$\frac{4(6-r)+V''v_l'v_k'}{V'v_k''-V''(v_k')^2} < -1 \tag{A8}$$

$$4(6-r) + V''v_l'v_k' < V''(v_k')^2 - V'v_k''$$

$$4(6-r) + V'v_k'' < V''(v_k')^2 - V''v_l'v_k'$$

$$4(6-r) + V'v_k'' < (v_k' - v_l')V''v_k'$$

$$\frac{4(6-r)+V'v_k''}{V''v_k'} > (v_k' - v_l')$$

$$v_l' > v_k' - \frac{4(6-r)+V'v_k''}{V''v_k'} > v_k'$$

So it is necessary that $v_l' > v_k'$ for condition (A8) to hold.

The utility function

$$U = 4(r + (6-r)l)k + V[M_0 - v_l(l) - v_k(k)]$$

Has the following second-order conditions

$$\frac{\partial U}{\partial k} = 4(r + (6-r)l) - V'[M_0 - v_l(l) - v_k(k)]v'(k)$$

$$\frac{\partial^2 U}{\partial k^2} = V''[M_0 - v_l(l) - v_k(k)]\big(v'(k)\big)^2 - V'[M_0 - v_l(l) - v_k(k)]v''(k)$$

$$\frac{\partial^2 U}{\partial k^2} = V''(v_k')^2 - V'v_k''$$

$$\frac{\partial^2 U}{\partial k \partial l} = 4(6-r) + V''[M_0 - v_l(l) - v_k(k)]v'(k)v'(l)$$

$$\frac{\partial^2 U}{\partial k \partial l} = 4(r-6) + V''v_k'v_l'$$

$$\frac{\partial U}{\partial l} = 4(6-r)k - V'[M_0 - v_l(l) - v_k(k)]v_l'(l)$$

$$\frac{\partial^2 U}{\partial l^2} = \big(V''[M_0 - v_l(l) - v_k(k)]v_l'(l)^2 - V'[M_0 - v_l(l) - v_k(k)]v_l''(l)\big)$$

$$\frac{\partial^2 U}{\partial l^2} = V''(v_l')^2 - V'v_l''$$

Hessian matrix

$$H = \begin{bmatrix} \dfrac{\partial^2 U}{\partial l^2} & \dfrac{\partial^2 U}{\partial k \partial l} \\ \dfrac{\partial^2 U}{\partial k \partial l} & \dfrac{\partial^2 U}{\partial k^2} \end{bmatrix} = \begin{bmatrix} V''(v_l')^2 - V'v_l'' & 4(6-r) + V''v_k'v_l' \\ 4(6-r) + V''v_k'v_l' & V''(v_k')^2 - V'v_k'' \end{bmatrix}$$

The Maximum condition

$$(V''(v_l')^2 - V'v_l'')(V''(v_k')^2 - V'v_k'') - (4(6-r) + V''v_k'v_l')^2 > 0$$

$$\big({V''}^2(v_l')^2(v_k')^2\big) - V''(v_l')^2V'v_k'' - V'v_l''V''(v_k'')^2 + {V'}^2v_l''v_k'' - A^2 - 2AV''v_k'v_l' - {V''}^2{v_l'}^2{v_k'}^2 > 0$$

$$-V''(v_l')^2V'v_k'' - V'v_l''V''(v_k'')^2 + {V'}^2v_l''v_k'' - A^2 - 2AV''v_k'v_l' > 0$$

$${V'}^2v_l''v_k'' - A^2 > V''(2Av_k'v_l' + (v_l')^2V'v_k'' + V'v_l''(v_k'')^2)$$

$$\frac{(V''(v_l')^2 - V'v_l'')}{(4(6-r) + V''v_k'v_l')}\frac{(V''(v_k')^2 - V'v_k'')}{4(6-r) + V''v_k'v_l'} > 1$$

On the SOC can be written

$$(V''(v_l')^2 - V'v_l'')(V''(v_k')^2 - V'v_k'') > (4(6-r) + V''v_k'v_l')^2$$

Take absolute values. This has the form AB > C^2, implying that A > B iff A > B > C. Therefore

$$\frac{4(6-r)+V''v_l'v_k'}{V'v_k''-V''\left(v_k'\right)^2} = \frac{dk}{dl} < -1V'v_k'' - V''(v_k')^2 < V'v_l'' - V''(v_l')^2 \tag{3}$$

To establish this, it is sufficient that $v_l'(x) > v_k'(x)$ and $v_l''(x) > v_k''(x)$ for all *x*.

$$Y = 14(1+l)k(l)$$

$$Y = 14\big(k(l) + lk(l)\big)$$

$$\frac{dY}{dl} = 14\left(\frac{dk}{dl} + \left(l\frac{dk}{dl} + k(l)\right)\right)$$

$$\frac{dY}{dl} = 14\left(k + (1+l)\frac{dk}{dl}\right)$$

FOC

$$0 = \frac{dk}{dl}(1+l) + k(l)$$

$$0.598(1+l) = k$$

$$0.598l = k(l) - 0.598$$

$$l = \frac{k}{0.598} - 1$$

Given an "initial condition" of (*k*,*l*) = (0.731,0.555) and the slope of the function, this implies the curve is defined by $k = 0.731 + \frac{dk}{dl}(l - 0.555)$.

$$0.598(1+l) = 0.731 - 0.598(l - 0.555)$$

$$l = \frac{0.731 + 0.598 * 0.555 - 0.598}{(0.598 + 0.598)}$$

$$0.598(1 + l) = 0.619 - 0.598(l - 0.742)$$

$$l = \frac{0.619 + 0.598 * 0.742 - 0.598}{(0.598 + 0.598)}$$