Pre-Analysis Plan

The Distribution of 1st and 2nd Rolls in Lying Experiments: Distinguishing Between Competing Theories

Paul Clist*& Ying-Yi Hong

8th November 2018

1 Introduction

Two important papers make predictions about the distribution of lying behaviour in dice rolling experiments. Abeler et al. (2018, p.1) review 72 related studies (comprising 32,503 subjects), arguing that the only way of rationalising the results are to use a model that combines "a preference for being honest with a preference for being seen as honest". By contrast, Gächter and Schulz (2016) are influenced by psychological research. They focus on an experimental detail of Fischbacher and Föllmi-Heusi's (2013) original paradigm that has been implemented for more than 10,000 people: subjects are asked to roll the die twice but report only the first roll. Influenced by Shalvi et al. (2011), Gächter and Schulz (2016) argue that subjects will report the higher of their two rolls, as this feels less like lying.

Abeler et al. (2018) do not make specific predictions on the final distribution of lying behaviour, as their model has two free parameters. Gächter and Schulz (2016) do make clear predictions, for example that with a six-sided die we should expect the maximum roll to be reported in 11/36 cases. Current evidence, despite being voluminous, is unable to test core predictions of Gächter and Schulz's (2016) Justified Dishonesty (JD) benchmark because it only records the first (incentivised) roll. Our new experiment is the first we're aware of that records the second (unincentivised) roll. This allows direct tests of the 'other half' of JD's predictions, for example that the maximum roll should be reported as the second roll in only 1/36 cases. Further, we are able to discuss the accuracy of different features of the two models, as well as making clear the links between them.

2 Research Design

The experiment is simple, and follows the standard protocol of Fischbacher and Föllmi-Heusi (2013) and Gächter and Schulz (2016) as far as possible. Namely, it will be a computer-based experiment on Z-tree where subjects roll a die and are paid according to the number reported. Subjects will not be observed, and have an incentive to misreport as reports 1-5 attract that number of monetary units, while 6 attracts nothing. We will run the experiment after one other experiment at the University of East Anglia's CBESS lab. The requested sample size is 250, which typically means we expect a sample of 200-250. We will have no control over the data collection or sampling, and will not see any results until all data has been collected. Payouts will be the same as Gächter and Schulz's (2016) UK sample, with numbers 1-5 attracting that number of monetary units (50p) and 6 attracting nothing. Selection is via a general email to CBESS's pool of participants, with subjects accepting on a first-come-first-served basis.

The original design had the following wording, with each number representing a new screen. Throughout a payoff table was present.

*https://paulclist.github.io

1. For the following questionnaire you will receive a small additional payoff. However, this payoff is not the same for every participant. You determine your own payoff by throwing your die twice as soon as you are asked to.
   Your first throw decides on how much you receive. You can see the exact payoff from the following chart. It will remain on the screen until you have entered your throw.
   The second throw only serves to make sure that the die is working properly. You may of course throw the die more than twice. However, only the first throw counts.
   If you have any question, please raise your hand. If you are ready, please press OK

2. Please throw the die now.
   Please keep in mind the first number you have thrown.
   If you have thrown the die, please press OK.

3. Now please enter the number i.e. the first number you have thrown.
   Number thrown: (with box to enter)
   Resulting payoff: (with box that updates)

Our new experiment has two treatments. Treatment A is below, with changes from the original Fischbacher and Föllmi-Heusi (2013) design in bold. We try and maintain the same visual presentation. As the first screen is identical, we start at the second:

2. Please throw the die **twice** now.
   Please keep in mind the first **and second numbers** you have thrown.
   If you have thrown the die **twice**, please press OK.

3. Now please enter the **first number you have thrown**.
   1st number thrown: (with box to enter)
   Resulting payoff: (with box that updates)

4. *Without payoff table*
   Now please enter the second number you have thrown.
   2nd number thrown: (with box to enter)

Treatment B is below, with changes from the original Fischbacher and Föllmi-Heusi (2013) design in bold. As the first screen is identical, we start at the second:

2. Please throw the die **twice** now.
   Please keep in mind the first **and second numbers** you have thrown.
   If you have thrown the die **twice**, please press OK.

3. *Two boxes, side by side. Box 1:*
   Now please enter the first number you have thrown.
   1st number thrown: (with box to enter)
   Resulting payoff: (with box that updates)

   Box 2, on the right
   Now please enter the second number you have thrown.
   2nd number thrown: (with box to enter)

Our interest is in testing the Justified Dishonesty Benchmark against both first and second roll behaviour. This presents a problem: there is no way to follow Fischbacher and Föllmi-Heusi's (2013) protocols exactly while also recording a second roll. In the original instructions subjects only read they should roll a die (at least) twice on screen one, but it is not mentioned again. This may mean that many subjects who were instructed to roll twice simply didn't. If we use a design that forces students to

report twice, that may mean a greater tendency to actually roll twice: potentially suggesting subjects follow a Justified Dishonesty strategy in our design when they would not have done so in the original.

Our solution to this is to use two treatments. A priori, both are good tests of JD. If it accurately describes behaviour, it should do so in both treatments and both rolls. Evidence also rests upon more than one design.

There are differences between the two treatments which may lead to different behaviour, but we do not have strong prior expectations, especially over the first roll reported. Treatment A is closest to the original design for the first 3 screens: we add only five words, delete three and pluralise one. The request to report the second roll only occurs after the first roll has already been reported, and so we believe the first roll in treatment A would be identical to what would have occurred in the original design.

Treatment B is a larger departure in the visual design on the third screen (two boxes are used) and changes the first decision into a concurrent one. Any difference in the reported first roll between treatment A and B *implies* subjects are not following a JDB strategy, as knowledge of the second roll being recorded should not influence the first roll. Any difference between the two treatments would be consistent with the idea that in the original design subjects do not actually tend to roll the die twice. While they are instructed to do so, they are not reminded to do so. If some subjects arrive at the third screen only having rolled once (or perhaps not at all) then a justified dishonesty strategy is not available to them in treatment A. Treatment B by contrast could *cause* JD behaviour, by almost suggesting subjects switch the two numbers. Hence we think that treatment B is a little more likely to trigger a justified dishonesty strategy.

It is worth noting briefly a difference between the original design and ours. Subjects are asked to remember 2 rolls rather than 1, which if anything is more likely to induce JD behaviour than the original design. Humans are generally thought to be able to remember 7 numbers, so remembering 2 doesn't appear to be a taxing requirement.

To conclude, we have chosen a design which is slightly in favour of JD, potentially pushing subjects slightly into a more JD-influenced direction. As such, it is a strong test of JD: if the theory does not fit behaviour here, it is not an accurate description of behaviour.

3 Hypotheses

Using Abeler et al.'s (2018) notation, $\bar{r}$ means the average report and $F(r)$ the distribution of reports. These are measured as the number of monetary units claimed, so lie between 0 and 5. We use subscripts 1 and 2 to denote the two rolls, and A and B to denote the two treatments. The 5 null hypotheses of interest are:

1. H_{JDB1} : The Justified Dishonesty Benchmark describes behaviour for the incentivised roll. This can be measured in two ways:
   - $\bar{r}_1 = 3.47$
   - $F(r_1) = 1/36 + (2/36) \cdot r$
2. H_{FHB1} : The Full Honesty Benchmark describes behaviour for the incentivised roll. This can be measured in two ways:
   - $\bar{r}_1 = 2.5$
   - $F(r_1) = 1/6$
3. H_{JDB2} : The Justified Dishonesty Benchmark describes behaviour for the unincentivised roll. This can be measured in two ways:
   - $\bar{r}_2 = 1.53$
   - $F(r_2) = 1/36 + (2/36) \cdot (5 - r)$

4. H_{FHB2} : The Full Honesty Benchmark describes behaviour for the unincentivised roll. This can be measured in two ways:
   - $\bar{r}_2 = 2.5$
   - $F(r_2) = 1/6$
5. H_{AB} : There is no difference in behaviour by treatment. This can be measured in four ways:
   - $\bar{r}_{1A} = \bar{r}_{1B}$
   - $\bar{r}_{2A} = \bar{r}_{2B}$
   - $F(r_{1A}) = F(r_{1B})$
   - $F(r_{2A}) = F(r_{2B})$

4 Empirical Strategy, with code

Where possible, we follow Gächter and Schulz's (2016) specific tests, using T tests for means and the KSD exact test for distributions[1]. We depart by making use of multiple testing corrections, given that there is more than one way of testing a given hypothesis. In order to aid comparison with previous work, we will still report the uncorrected p values, alongside the corrected q values.

The first tests we will perform relate to H_{AB}. In the below Stata code, `roll1` and `roll2` refer to the reported first and second roll respectively, with `treatment` capturing the two treatments. Given there are 4 relevant tests, rejecting equality on the basis of any one test would lead to over rejection of true null hypotheses. Specifically, if there was truly no effect we would reject at the 5% level in 18.5% of cases, rather than at 5%. As such, we will use the Hochberg step-up procedure to control the family-wise error rate. Rejection at the 5% level in any of these four tests would then lead to a rejection of H_{AB}.

```
    **** tests of treatment equality ****
gen pfamily1 = . // 1st family of tests: does A=B
ttest roll1, by(treatment)  // mean, roll 1
  replace pfamily1=r(p) if _n==1
escftest roll1, group(treatment)  // distribution, roll 1
  replace pfamily1=r(p_val) if _n==2
ttest roll2, by(treatment)  // mean, roll 2
  replace pfamily1=r(p) if _n==3
escftest roll2, group(treatment)  // distribution, roll 2
  replace pfamily1=r(p_val) if _n==4
qqvalue pfamily1, method(hochberg )  qvalue(qfamily1)
  list pfamily1 qfamily1 if qfamily1!=.
```

If we reject H_{AB} for any of the tests at the 5% level (using Hochberg q values), we will run all subsequent tests on both treatments individually. When reporting q values, these will then be corrected for 4 tests (treatments A and B, on the mean and distribution). If we do not reject H_{AB}, we will pool the two treatments and report q values merely corrected for 2 tests (mean and distribution).

For ease of comprehension and brevity, we will display the code below as if we can pool the data from the two treatments. We will also leave out the multiple testing correction, given this can clearly be seen in the code above. If the data can be pooled, this then gives four hypothesis tested, each by two tests (with q values corrected for 2 tests). If data cannot be pooled, this gives four hypothesis tested, each by four tests (with q values corrected for four tests).

[1] When testing for equality of distributions between A and B, we cannot use the KSD test and so opt for the Epps-Singleton test.

```
    **** JDB, roll 1 ****
ttest roll1==3.47 // test of mean, JDB
mgof  roll1 = (2/36)*roll1 + 1/36, mc ksmirnov // distribution, JDB

    **** FHB, roll 1 ****
ttest roll1==2.5 // test of mean, FHB
mgof  roll1 = 1/6, mc ksmirnov // distribution, FHB

    **** JDB, roll 2 ****
ttest roll2==1.53 // test of mean, JDB
mgof  roll2 = (2/36)*(5-roll2) + 1/36, mc ksmirnov // distribution, JDB

    **** FHB, roll 2 ****
ttest roll2==2.5 // test of mean, FHB
mgof  roll2 = 1/6 , mc ksmirnov // test of distribution, FHB
```

5 Test of the ANR model

We cannot test Abeler et al.'s (2018) model in the same way as the above, as they do not make a single set of predictions. Rather they include two free parameters which mean that certain behaviour can be rationalised by the model, and certain behaviour cannot. Our proposed test of the ANR model is then to see whether the final distributions we find can be rationalised by the proposed model. The 'test' of their proposed model is simply whether there exists a set of $[k_1, k_2]$ that rationalise the final distributions.

Note that despite the similarities between the two models, the final distribution of reports predicted by Gächter and Schulz (2016) is not compatible with the model proposed by Abeler et al. (2018). In other words, even the first roll evidence will be able to distinguish between the models. Of the common features, one crucial element is that both models predict that an equal proportion of subjects receiving any true state ω will report any given $r \neq \omega$. This is clear from the use of linear lying costs, with the reputation cost of being seen as a liar equally felt by subjects whatever their true state. (Note that at the time of writing only the working paper version is available.)

6 Relation to Previous Experiment

This study is in part motivated by a previous experiment, which we'll refer to here as Experiment 1. That was the first 2 questions in a stand-alone paper and pen survey, using double the pay outs with a £2 show-up fee and 4 subject pools. In terms of the design itself, subjects were asked to roll and report the first dice as question 1, before being asked to roll and report a second time as question 2. Results strongly conformed to the JDB in each of the four subgroups for question 1, despite the fact that technically the mechanism *shouldn't* have been at play. The second roll was very close to the full honest benchmark. We expect to use the same tests in the presentation of the 1st experiment, namely:

1. A test of mean (two-sided t-test) and distribution (KSD test)
2. against the JDB and full honesty predictions
3. for first and second roll
4. for each of the four sub groups

The corresponding Stata code is below. Note we do not include this in order to claim the below analysis was preregistered. Rather, we wish to tie our hands at this point so as to not have researcher degrees of freedom to reanalysis experiment 1 based on the outcome of experiment 2 (at least without noting this is exploratory).

```
bysort treatment:  ttest roll1==3.47
bysort treatment:  ttest roll1==2.5
bysort treatment:  mgof roll1 = (2/36)*roll1 + 1/36 , mc ksmirnov
bysort treatment:  mgof roll1 = 1/6 , mc ksmirnov

bysort treatment:  ttest roll2==1.53
bysort treatment:  ttest roll2==2.5
bysort treatment:  mgof roll2 = (2/36)*(5-roll2) + 1/36 , mc ksmirnov
bysort treatment:  mgof roll2 = 1/6, mc ksmirnov
```

The differences between experiment 1 and experiment 2's treatments A and B will not be tested explicitly, rather they will all be tested against the theoretical predictions. Experiment 1 has points of similarity with each of our treatments. Like treatment A, there is a sequential report of the two rolls. Like treatment B, there is the chance (if a subject reads ahead, or changes a previous answer) to consider the two reports at the same time.

References

Abeler, J., Nosenzo, D., and Raymond, C. (2018). Preferences for truth-telling. *Quarterly Journal of Economics*.

Fischbacher, U. and Föllmi-Heusi, F. (2013). Lies in disguise - an experimental study on cheating. *Journal of the European Economic Association*, 11(3):525–547.

Gächter, S. and Schulz, J. F. (2016). Intrinsic honesty and the prevalence of rule violations across societies. *Nature*, 531(7595):496.

Shalvi, S., Dana, J., Handgraaf, M. J., and De Dreu, C. K. (2011). Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior. *Organizational Behavior and Human Decision Processes*, 115(2):181–190.