Promoting Academic Honesty: a Bayesian Causal Analysis of an Integrity Pilot Campaign[*]

Alejandro Puerta[†]and Andrés Ramírez-Hassan[‡]

We examine the effect of an integrity pilot campaign on undergraduates' behavior. As many costly small scale experiments and pilot programs, our statistical inference has to rely on a small sample size. To tackle this issue, we perform Bayesian retrospective power analysis. In our setup, a lecturer intentionally makes mistakes that favors students' grades, who decide whether to disclose them or not. We find evidence that at least in the short term, the pilot campaign has a positive impact on the students' disclosure probability.

Keywords: Academic integrity, Bayesian Inference, Honesty, Statistical Power

I. Introduction

Academic dishonesty has been widespread among college students for a long time. Bowers (1964) found that 75% of the students in his research have been engaged in one or more incidents of academic dishonesty. Haines et al. (1986) assert that more than 50% of the interviewed students reported cheating. However, there is no much empirical research about incentives to cheat (Martinelli et al., 2018), despite that McCabe, Treviño and Butterfield (2001) find that such dishonest behavior is prevalent. In fact, Grimes and Rezek (2005) discover that the majority of high school economics students in six transitional economies and the USA had cheated. Even though reported cheating incidence varies greatly across studies (see McCabe and Trevino (1997); Park (2003)), dishonest behavior in the academy seems to be prevalent. Davis et al. (1992) assert that even though more than 90% of students (among more than 6,000 respondents) answered that cheating is wrong, 76% of them reported that they have cheated in an academic environment. So, even though students bare in mind that they should not cheat, they do. As a response, some business schools are taking ethics into account in their curricula and giving weight to ethical students' orientation in admission processes (McCabe, Butterfield and Trevino, 2006).

The findings from McCabe, Treviño and Butterfield (2001) suggest that "cheating can be most effectively addressed at the institutional level". However, imple-

[‡] Universidad EAFIT. Medellín, Colombia. Email: aramir21@eafit.edu.co. **Corresponding author**.

[*] We thank the former chancellor of Universidad EAFIT, Dr. Juan Luis Mejía Arango, and the head of the Center of Integrity at Universidad EAFIT, Nathalia Franco Pérez, for your support and guidelines in this project.

[†] Department of Economics, Universidad Carlos III, Madrid, Spain, E-mail: alpuerta@eco.uc3m.es

mentation of large scale campaigns requires considerable costs, which is why pilot programs are carried out, and statistical inference is commonly performed to decide whether to incur or not in such sunk costs. Due to budget restrictions, such pilot programs involve a frequently small sample size that comes along with statistical power concerns. Our contribution to the literature is to perform a Bayesian analysis to assess the impact of an integrity pilot campaign on undergraduate students' academic integrity. Our econometric framework allows to tackle the sample size issue and provides an easy to implement framework in similar settings, such as small scale experiments, to perform power analysis.

Once a pilot campaign or experiment has been carried out, the sample size is fixed and the practitioner must cope with it. This implies that efforts must be made to perform estimation as robust as possible given the limited data. Accordingly, we propose to perform a retrospective power analysis. From a Bayesian perspective "the power of an experiment is the probability of rejecting the null hypothesis if the data were generated from a particular specific alternative effect size" (Kruschke, 2013). We implement a retrospective power analysis that consists on three steps: (i) estimating the model (ii) simulating counterfactuals based on the posterior chains and (iii) calculating the proportion of times the results lend evidence against a given set of hypotheses (given that they are false). This is particularly powerful if the objective is to assess the effect of a key independent variable in the presence of a small sample size. The reason is that based on the characteristics of the observed individuals, counterfactuals are simulated for the dependent variable, given different effect sizes.[1] Consequently, given that only one realization of the dependent variable is known for each observation, performing retrospective power analysis lends very useful information. In particular, in the light of small sample sizes.

Our study consists of identifying the conditional causal effects of interventions that promote academic integrity. Such interventions correspond to a pilot program carried out by the Center of Integrity of a college in Medellín, Colombia. They designed a structured observation[2] in which a lecturer intentionally makes mistakes that favors students' grades. We attempt to assess the incidence of attending the integrity interventions on the students' decision of disclosing or not such mistakes. The experiment is designed so that the students realized the lecturer's mistake, and faced the decision about whether to tell him to correct their grade. 36 students were observed sequentially to obtain a longitudinal data-set with a sample size equal to 88. Students were equally and randomly split into two groups: treatment and control, where the former group is exposed to the interventions. In this framework, the external costs are null, since the students keep their quizzes, so the probability that they are "caught" is zero. Additionally, the design promotes a high level of displacement of own-responsibility, so that the

[1] Different effect sizes refer to the values of the chains obtained in the Bayesian estimation, which implies that they are consistent with the observed data.

[2] This means a direct observation in a natural setting, where researchers intervene to have some control over the observed events (Shaughnessy and Zechmeister, 2012)

probability that individuals judge themselves as dishonest is low when they are in fact being dishonest (Bandura, 2002).

The characteristics of our setup contribute to an accurate identification of the causal effect. First, the “structured observation” seems to mitigate the self-selection and behavioral modifications associated with self-reporting surveys. Such an experimental setting is well known in psychology (Hyman et al., 2010; Valentino et al., 2011; Piaget, 2013), but, to the best of our knowledge, it is not common in economics. Second, we control for psychological, educational, habit-related, and socioeconomic variables, which allows us to mitigate the effect of confounding variables. Third, we use a Bayesian approach which is exact and does not rely on large samples (Greenberg, 2008), this is particularly desirable in presence of a small sample size. Since we estimate a hierarchical random–effects model, we take into account individual unobserved heterogeneity in unbalanced panel settings. Most importantly, we perform an easy to implement algorithm to perform retrospective Bayesian power analysis to check the robustness of our results in the light of the small sample size.

Although the integrity interventions had a moderate impact, the results from our experiment are hopeful for a society such as the Colombian, which suffers from high levels of corruption, according to the corruption perceptions index (CPI) (Index, 2019). Even though students face a very tough environment for being honest, we found that attendance of an additional integrity intervention session increases the probability of disclosing the lecturer’s mistake on average by 3.95 points, with a 95% credible interval of (2.84, 4.87). The power test of not having a significant effect is 0.98 with a 95% credible interval equal to (0.97, 0.99). This evidence suggests that, at least in the short term, the availability of moral standards pushes internal rewards up, and as a consequence, the treated economics students have a higher probability of acting honestly.

The remainder of this paper proceeds as follows. Section II describes the experiment and the integrity interventions. Section III describes the econometric methodology. Our main results are described in Section IV. Section V concludes.

II. Design: experiment and interventions

A. *Context*

As Grimes (2004) suggests, personal attitudes and ethical standards can be greatly influenced by academic standards, cultural differences, religious beliefs, and the political regime, which are commonly country or region-specific. This makes relevant to broadly describe the context where the experiment was carried out. It was conducted at a university in Medellín, Colombia. In Latin America, violations of academic integrity appear to have been more frequent over the past twenty years, which could be due to the high levels of perceived corruption in the public sector (García-Villegas, Franco-Pérez and Cortés, 2016). Moreover, Colombia occupies the 99th position out of 180 countries in terms of the CPI (Index, 2019). However, Colombia is 7th in terms of the proportion of Catholics in its population (Pontificio, 2014). In particular, around 80% of the Colombian population is declaredly Catholic. So, although Colombia is a very religious country, there is recent evidence of high incidence rates of fraud and violation of academic integrity (García-Villegas, Franco-Pérez and Cortés, 2016). At the regional level, the culture of the population of Medellín is enterprising and creative, but being sly is perceived as a good personal feature. Evidence for this is that 96% of a representative sample of undergraduate students in this university acknowledged having committed fraud at least once (García-Villegas, Franco-Pérez and Cortés, 2016).

B. Setup

Our experiment involves undergraduate economics students attending an Econometrics course. The average grade in this subject (3.7/5.0) is lower than the average economics GPA (3.9/5.0), which could make the decision to disclose harder. Econometrics students are always split into two groups, due to infrastructural limitations in the computer labs. One of them was chosen at random to be the treatment and the other the control group. There were 36 students, equally split into each group, we randomly selected 4 students during 12 classes from each group, at the end we have 88 observations.[3] Both groups have the same lecturer, who lectured two classes (theory and applications) every week for four months. These groups are joined in the weekly theory lecture, and split for the computer (applied) class. In the latter, the treatment group was exposed every week for three months to interventions (twelve in all) that promote integrity. These interventions were carried out as a pilot program due to university planning to implement an integrity campaign at the institutional level. The students were informed of this fact at the end of the experiment. Each intervention lasted 15 minutes, and was made at the end of the class, to guarantee the students' attendance; the class lasted 90 minutes, including the intervention. The net time for the applied class was 75 minutes in both groups.

C. Methodological aspects

The identification strategy for the causal effect was based on the grading system of the course. In particular, 60% of the total grade was based on 12 quizzes, given each week for three months. Each quiz had ten multiple choice questions, each correct answer worth 0.5 points, so that a perfect score is equal to 5.0. Each quiz had the same weight (5%). All students were evaluated at the same time with

[3] Notice that the potential number of observations is 96 (4 students × 2 groups × 12 quizzes). Nonetheless, three students (two from the control group and one from the treatment group) did not answer the survey, which was highly encouraged to answer but not compulsory. For this reason, they could not be taken into account for the final sample size.

the same quizzes, hence the control and treatment groups faced exactly the same grading conditions.

To identify the causal effect, at each quiz, the lecturer randomly selected four students from each group and gave them 0.5 extra points (one right answer) apparently by mistake. To reinforce the impression that this was a mistake, and not a bonus, all the solutions were explained at the moment of handing back the marked quizzes, and the right number of correct answers was written next to the wrong grade (see Figure 5 in the Appendix). Since there is no guarantee that students consciously recognized that a mistake had been made, it is likely to have measurement error in the data. However, such measurement error is random between treatment and control groups. In addition, it is in the dependent variable, so it only results in a larger error variance, but the estimators remain consistent and the inference procedures remain valid (Wooldridge, 2010).

Given this setup, our dependent variable is *disclosing* ($y_{it} = 1$) or *not disclosing* ($y_{it} = 0$) to the lecturer his grading mistake. Students had the opportunity to do so throughout the semester. The exams were delivered in class, so all the students that disclossed did it by means of approaching the lecturer after class. Some of them (those who disclossed) did it the next day and others took more time, but there was no single "public disclosure". Our design tried to minimize the possibility that "lucky" students did not notice the lecturer's mistake. However, if this issue is present, it makes sense to assume that it is randomly distributed between all the students, because the control and treatment groups were exposed to exactly the same design, except by the interventions. Observe that the structured observation (Shaughnessy and Zechmeister, 2012), guards against the self-selection and behavioral modification associated with self-reporting surveys. These are desirable characteristics, which help for a more accurate identification of the causal effects associated with the moral integrity interventions.

Since the lecturer handed back the graded quizzes in our experiment and the students kept them, it was supposed by them that the lecturer did not have any evidence to corroborate his mistake. So, the probability that students are "detected" being dishonest is zero, and as a consequence, there is no external punishment. In addition, the experiment is based on the lecturer's mistakes. This implies a high probability that students displace their responsibility for *not disclosing*, so that it would not be considered dishonest (Bandura, 2002). However, it is a dishonest action because the grade does not correspond to what the student deserves. This is a very tough environment in which to be honest, which in turn is based entirely on internal reward. Accordingly, a set of interventions were designed based on activities which have shown to elicitate honest behavior in the literature. Their main objective is to promote academic integrity among students, representing an effort to enhance the internal reward associated with *disclosing* relative to the external reward of *not disclosing*.

First, students were asked to voluntarily answer a survey to collect living habits, as well as socioeconomic, educational and psychological characteristics. Then, the treatment group agreed to attend the integrity interventions, and at the same time, the experiment was performed. After it was over, all students were informed about having been part of this experiment, and agreed to participate in a focus group to provide feedback and qualitative insights into the experiment; they also agreed to the use of these results for academic purposes.

In our experiment, students did not know that their behavior was being monitored. We proceeded in this way so as to measure the individuals' behavior as accurately as possible, aiming to mitigate the self-selection and behavioral modifications coming from self-report surveys or self-selection to participate. In any case, students' anonymity was preserved.

As Ortmann and Hertwig (2002) assert, the "plausibility" of deception highly depends on its severity, methods of debriefing and the recruitment mode. Students agreed to participate in the integrity interventions and they answered a survey with their socioeconomic, educational and living habits. Students were debriefed at the end of the semester.

There are two main issues arising when deceiving: methodological (Bonetti, 1998), because it could make the data be invalid, and ethical, in the sense that we could harm others. Regarding the latter, Kelman (1967) suggests that the primary way of counteracting negative effects is post-experimental feedback. So, we informed students at the end of the experiment, and asked for their approval. They liked the experiment, and its main objective. Then, they agreed to participate in a focus group to obtain qualitative insights, and that we use the outcomes for designing a moral integrity campaign, and publishing the experimental outcomes (always maintaining students' anonymity).

Observe that "...costs and benefits matter, so that such practices (deception) might in fact be appropriate when the topic is important and there is no other way to gather data" G. Charness and van De Ven (2021). We consider that this was our case: it is very difficult to get data without altering students' behavior in our situation, and learning what promotes integrity in the academy is particularly worthy in a society like ours. On the other hand, it seems that undergraduate students who had participated in experiments bothered about deception when this implies more work and/or affects their pay, this was not our case, and less upset when there is an omission of information, our case, although we reveal all information at the end.

Notice that our experimental design always benefits the students with a higher grade. According to what students shared in the focus group, there were apparently no negative consequences for them.

D. Interventions

The objective of the interventions is to make moral standards more available, to push the internal reward up, and promote honest behaviour in undergraduate economics students. As mentioned above, they were based on previous psychological and economical experiments showing positive effects on students' integrity, namely: lectures about academic integrity, discussions about this issue, documentaries related to integrity, and ethical dilemmas, among others. Table 1 shows a brief description of the interventions, which elicitation each one sought, and the evidence which motivated them.[4]

To give insight on the interventions in Table 1, we briefly describe how they were carried out. Mazar, Amir and Ariely (2008) assert that "making people mindful by increasing their attention to their honesty standards can curb dishonesty." Accordingly, several experiments were carried out, looking for an increase of the attention to standards of honesty. For example, students were given a bracelet saying "Dare to think" which is the slogan of the university's integrity campaign (Week 6). Mazar, Amir and Ariely (2008) suggest that remembering the Ten Commandments prompts honest behaviour whether or not people have religious attachments; so, quizzes were given to the treatment group with the Ten Commandments as a heading (Week 1). McCabe and Trevino (1997) found evidence suggesting honor codes promote integrity; therefore, students were asked to sign an honesty commitment (Week 5). Jordan, Mullen and Murnighan (2011) found evidence suggesting that when people remember an immoral action, they tend to act more honestly in order to "compensate". So students were asked to share stories about dishonest acts (Week 12). See Appendix VI.B for details.

The integrity interventions for the experiment were done in the framework of structured observation. In this set-up, researchers attempt to keep the environment as natural as possible to avoid as much as possible that results are distorted. Nonetheless, they intervene to exert some control over the events they are observing. For this reason is not uncommon that the individuals know they are being observed ex-post, as in Hyman et al. (2010). There are, however, potential issues that could arise with the use of this methodology, which are outlined by Shaughnessy and Zechmeister (2012): (i) when researchers fail to follow similar procedures whenever they make an observation, is less likely that the obtained results would be the same when investigating the same problem, (ii) unknown variables could play an important role in producing the behavior under observation, (iii) there could be intentional blindness. That is, people fail to notice new and distinctive stimuli in their environment. Regarding the first two issues, Shaughnessy and Zechmeister (2012) suggest that to prevent such problems, researchers must be consistent in their procedures and "structure" observations as similarly as possible across observations. For this reason, instead of evaluating with midterms and finals, which could have different structures, the mistakes were made in 12 quizzes based on multiple choice answers with the same number of questions and same weight in the final grade. In addition, we control for potential confounders. Regarding the third issue, we attempt to be as evident as possible with the mistake, placing the right number of correct answers next to the wrong

[4] To design the interventions, a literature review was carried on, searching for which kind of actions, activities or experiences elicited acting honestly. The references in Table 1 point out to the article that used as a basis for designing the corresponding interventions.

grade (see Figure 5 in the Appendix).

Table 1—: Integrity interventions

Week	**Intervention**	**Elicitation**	**Reference**
1	Lecture: Ethics and academic integrity and ten commandments in quiz	Reminder of academic integrity	Mazar, Amir and Ariely (2008)
2	Fragment of the documentary "Inside Job"	Ethical dissonance	Barkan, Ayal and Ariely (2015)
3	Lecture: Integrity in the professional environment	Consequences of dishonest actions	Bandura et al. (1996)
4	Ask explicitly if the grade is OK and ask about satisfaction in class	Obligation to lie and satisfaction needs	Gneezy (2005) and Ryan and Deci (2000)
5	Institutional campaign: *Atreverse a Pensar* and signing an honesty commitment	Euphemistic language and honor code	Bandura et al. (1996) and McCabe and Trevino (1997)
6	Gift: bracelet saying "Dare to think" and public recognition to a student who *disclossed* the lecturer's mistake in a specific class	Reminder of academic integrity and turning internal reward into external	Mazar, Amir and Ariely (2008) and Ayal et al. (2015)
7	Fragment of the series: "Brain games (Moral dilemmas)"	Reminder of integrity	Thaler (1980)
8	Lecturer delivers quizzes in his office, and obtains feedback from students about the class	Satisfaction of need	Ryan and Deci (2000)
9	Lecture: *Bancolombia* case. (a well-known event of corporate dishonesty)	Ethical dissonance	Barkan, Ayal and Ariely (2015)
10	Brain games episode: "Trolley Problem and Kin Selection"	Reminder of integrity	Thaler (1980)
11	Real testimony concerning cheating behavior	Consequences of dishonest actions	Bandura et al. (1996)
12	Classmates' stories about dishonest acts	Licensing and compensation	Jordan, Mullen and Murnighan (2011)

Note: Integrity interventions are based on literature review that shows that these interventions may have promote integrity.

III. Methodology

A. *Specification*

López-Pérez and Spiegelman (2013) suggest that the motivation for people to be honest could be due to pure lie aversion. Since we are interested in the conditional causal effect of the moral integrity interventions, the model should take into account whether the sequence of an individual's decisions is due to their characteristics rather than due to the interventions. Consistently, as some students were faced with the decision (disclosing/not disclosing) multiple times, which results in a non-independent set of observations, we implemented a longitudinal random effects normal model.[5] This approach allows us to take into account unobserved individual characteristics. Moreover, the reason for a student to disclose could be due to psychological characteristics that could be related to their educational, socioeconomic, or living habits features. Therefore, it is also convenient to control for psychological variables. Consistently, we conducted a survey including psychological, socioeconomic, educational, and living habits variables. This allows us to mitigate the effect of confounding variables, when aiming to identify the conditional causal effect of the integrity interventions.

One concern when identifying the causal effect of the integrity interventions is students' motivation to disclose. That is, students might not disclose the mistake because they are too shy to approach the lecturer, some of them might not do it because they are not very concerned with grades at all, among others. Consistently, we try to capture some of the motivational aspects of disclosing, asking students questions regarding their motivation to do well in class. In particular, we asked from 0 to 9 how much did they agree with a set of assessments, such as: (i) I study mainly because I like to learn, (ii) I value more knowledge acquisition than grades, (iii) I think is more important to know where I was wrong in an exam than the grade itself. These questions are summarised by a variable which we control for named *intrinsic motivation*.

A moral dilemma was set where students had to tell how much they agreed with several assertions. Then, we calculated the C-index, which measures the degree to which individuals let their judgment of behavior be determined by moral concerns or principles, rather than by other psychological forces (Lind, 2007). The calculation is based on the Moral Judgment Test, which is an instrument to measure an individual's moral-judgment competence as well as assessing their moral attitudes. This means measuring the ability of individuals to judge arguments in controversial moral problems on the basis of their own moral principles (Lind, 2008). The construction of the index has two parts (j), each one having six stages (l) corresponding to each moral stage according to Kohlberg's hierarchy (Kohlberg and Hersh, 1977). The first part includes the judgments supporting an action, whereas the second part includes the judgments criticizing it. The C-index is calculated as follows,

$$\frac{SS_{stage}}{SS_{dev}} \times 100,$$

[5] The reason why we do not estimate a binomial model (e.g., probit or logit) is because we are interested on calculating the marginal effects for the representative student, rather than performing prediction. The marginal effects from the normal model akin those of the nonlinear models for the representative individual. Accordingly, and for the sake of parsimony, we pick the normal model. Table 8 lends evidence in regard of how similar are the estimates, regardless of the model.

where

$$SS_{stage} = \frac{1}{2}\sum_{l=1}^{6}\left(\sum_{j=1}^{2} x_{lj}\right)^2 - \left(\frac{1}{2}\frac{1}{6}\sum_{l=1}^{6}\sum_{j=1}^{2} x_{lj}\right)^2 \text{ and } SS_{dev} = \sum_{l=1}^{6}\sum_{j=1}^{2} x_{lj}^2 - \left(\frac{1}{2}\frac{1}{6}\sum_{l=1}^{6}\sum_{j=1}^{2} x_{lj}\right)^2,$$

x_{lj} is the student's answer to part j of stage l.[6]

This index ranges between 0 and 100. Higher values imply that the individual's judgment is based on moral concerns rather than other psychological issues. In addition, we control for a measure of cognition related to students' motivation, using a Likert scale associated with each set of questions. Anderman, Griesinger and Westerfield (1998) indicate that cheating behavior is associated with motivational orientations and the use of science deep-level strategies. Therefore, we also asked the students questions regarding personal intrinsic motivation.

Regarding the socioeconomic, educational and living habits variables, we control for variables that have been used in previous research, such as household income (Bowers, 1964), parents' education (Bowers, 1964; McCabe and Trevino, 1997; Grimes and Rezek, 2005; Khodaie, Moghadamzadeh and Salehi, 2011), age (Haines et al., 1986; Lipson and McGavern, 1993; McCabe and Trevino, 1997; Anderman and Midgley, 2004; Finn and Frone, 2004; Grimes and Rezek, 2005; Kanat-Maymon et al., 2015), gender (Bowers, 1964; Lipson and McGavern, 1993; McCabe and Trevino, 1997; Finn and Frone, 2004), religion (Grimes and Rezek, 2005), GPA (Bowers, 1964; Haines et al., 1986; Lipson and McGavern, 1993; McCabe and Trevino, 1997; Finn and Frone, 2004; Grimes and Rezek, 2005), and membership in representative college groups (Bowers, 1964; Haines et al., 1986; McCabe and Trevino, 1997). In addition, we control for other variables, such as having a scholarship, out of class study hours, and living habits (smokers and drinkers of alcohol).

B. Econometric approach

We adopt a Bayesian approach for three main reasons. Firstly, since in each quiz four students from each group were randomly selected, we should incorporate that the same student can be randomly selected many times by taking into account unobserved heterogeneity. The family of random effects model allows this specification, without assuming that the panel is either balanced or equidistant. Secondly, Bayesian inference does not rely on large samples (Greenberg, 2008). Considering that we are dealing with a small sample, a Bayesian approach is better suited in this sense. Finally, the retrospective power analysis that we implement is only possible due to the availability of the posterior chains obtained from a Bayesian estimation.

Accordingly, we estimate a hierarchical random effects normal model, which

[6] In fact, we have a scaled version of the C-index. The original scale for the calculations ranges from -4 to 4 (in steps of 1) and our questions range from 0 to 9 (also in steps of 1).

has the form (Martin, Quinn and Park, 2011)

$$y_{it} = \boldsymbol{x}'_{it}\boldsymbol{\beta} + b_i + \epsilon_{it}, \tag{1}$$

where individual $i = 1, 2, \ldots, n$ has $t = 1, 2, \ldots, n_i$ observations, y_{it} is 1 if the student i disclosed at time t, and $\boldsymbol{x}_{it}$ is a set of covariates. $b_i \sim \mathcal{N}(0, \sigma_b^2)$ captures individual unobserved heterogeneity, $\boldsymbol{\beta}$ is a k-dimensional vector of fixed effects parameters, and $\epsilon_{it} \sim \mathcal{N}(0, \sigma^2)$ are stochastic errors. Our key independent variable is the accumulated interventions the student has attended. Taking into account that students would skip some interventions, we measure the causal effects associated with the accumulated number of interventions that the student attended at the moment of the decission instead of just indicating if the student belongs to the treatment group or not. Although we performed robustness checks based on the latter.

We use standard conjugate priors, that is, $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\beta}_0, \boldsymbol{B}_0)$, $\sigma^2 \sim \mathcal{IG}(\alpha_0, 1/\delta_0)$ and $\sigma_b^2 \sim \mathcal{IG}(d_0, d_0^{-1} v_0)$.[7] Additionally, "noninformative" priors are set in our estimation. In particular, we centered the fixed effect parameters at zero, and used an over-disperse diagonal covariance matrix with elements equal to 1.0E6. In addition, we set $\alpha_0 = \delta_0 = 0.001$ (Spiegelhalter et al., 2003), $v_0 = 1$ and $d_0 = 0.1$ (Martin, Quinn and Park, 2011). Intuitively this means that a priori we assume that none of the regressors have an effect on the dependent variable, and that we allow those parameters to vary a lot to cover the parameter space. We conducted a Gibbs sampler given standard conditional posterior distributions.[8]

C. Bayesian Retrospective Power Analysis

In a frequentist framework, power, significance, sample size and population effect size are related, so that any one of them is a function of the other three. Hence, when the values of the other three are fixed, the fourth is determined (Cohen, 2013). As noted by Kruschke (2013), many authors (Gerard, Smith and Weerakkody, 1998; Hoenig and Heisey, 2001; Nakagawa and Foster, 2004; O'Keefe, 2007; Steidl, Hayes and Schauber, 1997; Sun, Pan and Wang, 2011; Thomas, 1997) assert that conditional on only using data from a single experiment, the retrospective power analysis yields no additional information beyond what is already implicit in the p value.

On the other hand, O'Keefe (2007) states that "after-the-fact power analyses can sometimes be a useful supplement to p values and confidence intervals, but only when based on population effect magnitudes of independent interest". For example, Thomas (1997) highlights a useful case of retrospective power analysis:

[7] $\mathcal{N}$ stands for a normal distribution, and $\mathcal{IG}$ for a inverse gamma distribution.

[8] The Gibbs sampler is a Markov chain Monte Carlo (MCMC) algorithm which consists in sampling from the conditional posterior distributions of the parameters recursively in order to obtain the marginal posterior distribution for each parameter. See the Appendix, Section VI.C for more detail on the Gibbs Sampler for our model.

when one is willing to determine whether the study meets a standard target or to compare such findings with different studies; even when checking if a biologically (or other kind) meaningful pattern is met. Moreover, such post-hoc power analysis can at least make explicit the probability of achieving various goals in the given experiment (Kruschke, 2013).

A general framework for Bayesian retrospective power analysis proceeds as follows: "at every step in the MCMC chain of the posterior, the analyst uses that step's parameter values to simulate new data, then does a Bayesian analysis of the simulated data, and then checks whether the desired goals are achieved (...) From the many simulations, the proportion of times that each goal is achieved is used to estimate the probability of achieving each goal" (Kruschke, 2013).[9] In a frequentist approach, it is difficult to perform such analysis. The reason is that one only has the point estimate for the parameters, and its asymptotic distribution most of the time, instead of the chains for the posterior distribution.

We implement the aforementioned procedure to estimate the probability that the interventions have an effect on the probability of disclosing greater than a range of values, specially if it is greater than 0. We provide the step by step in Algorithm A1, but briefly discuss it below. The dependent variable ($y_{it}^{(s)}$) is simulated for s=1,2,...,S, using: $\boldsymbol{x}_{it}$, $\boldsymbol{\beta}^{(s)}$ from the posterior draws, and simulate $\epsilon_{it}^{(s)} \sim \mathcal{N}\left(0, \sigma^{2(s)}\right)$ according to draws from $\sigma^{2(s)}$, and $b_i^{(s)} \sim \mathcal{N}\left(0, \sigma_b^{2(s)}\right)$.[10] In this sense, we simulate several counterfactuals: given the same characteristics for a set of observations, what would have happened if the realization of the parameters and the idiosyncratic shock were different? Once y_{it} is simulated, we estimate the model in Equation (1) and get the posterior chains for the parameters. From this, we get the marginal effect associated with the number of interventions. Finally, we calculate the highest density interval (HDI) and check if the 5% lowest bound is bigger than each of the predefined thresholds. The proportion of times that this is true gives as an estimate of the probability that the interventions have an effect greater than a set of values.

Such approach offers the advantage of explicitly weighting the draws according to the posterior distribution. In particular, we have distributional information of the alternative hypothesis associated with different size effects (Kruschke and Liddell, 2018). We do not have to fix a particular value of the alternative hypothesis, and we take into account uncertainty regarding the unknown parameters.

[9] The usefulness of such a power analysis is conditional on the fact that the simulated data imitates the actual data (Kruschke, 2014). However, this is an implicit assumption one makes when modeling in a parametric framework: that the real data generating process (DGP) is the one assumed.

[10] Since the response variable is binary, we simulate the latent variable according to equation (1) and draw the response variable from a Bernoulli distribution by means of a logistic link function. In particular, $y_{it} \sim Bernoulli(\theta_{it})$ such that $logit(\theta_{it}) = y_{it}^{*} = log(\frac{\theta_{it}}{1-\theta_{it}}) = \boldsymbol{x}_{it}'\boldsymbol{\beta} + b_i + \epsilon_{it}$. As mentioned before, assuming for estimation a normal DGP instead of a logistic one, provides consistent estimates and has no particular drawbacks if one is interested in the marginal effects for the representative agent.

IV. Results

Table 2 presents descriptive statistics and tests for the mean differences between the control and treatment groups. In general, we can see that there are no significant statistical differences between the control and treatment groups, except for the probability of disclosing, which is our main objective variable, and mother's education, which is higher for the control group. To asses if our sample is representative of undergraduate population of the university, we conducted four different tests for the available variables in a historical survey of students from the university. We only used data for regular students at the academic period that the experiment was conducted, which is a sample size of 7,450 students, with information for: gender, age, income and mother's education of the students. For the age variable we conducted a difference in mean Welch's test as we did in Table 2. Since the other variables are categorical, we conducted a χ^2 goodness

Table 2—: Descriptive statistics: Difference in mean Welch's test

Variable	Mean Control	Mean Treatment	t-statistic
Proportion of disclosing	0.09	0.27	-2.25*
Age	20.50	20.87	-0.72
Female	0.44	0.60	-0.89
Mother's education	17.44	15.13	2.03*
GPA	3.98	3.86	1.06
Scholarship	0.06	0.27	-1.53
College groups	0.38	0.40	-0.14
Smoker	0.06	0.20	-1.11
Alcohol consumer	0.62	0.53	0.40
Scaled C-index	78.17	75.23	0.55
Intrinsic motivation	22.31	21.40	0.85
Income 0[a]	0.00	0.13	-1.47
Income 1[b]	0.19	0.33	-0.90
Income 2[c]	0.38	0.33	0.23
Income 3[d]	0.44	0.20	1.42
Study Hours	5.88	5.33	0.76

* Difference is statistically significant at 5%
[a] Less than 632 USD monthly
[b] 632–1580 USD monthly
[c] 1580–3160 USD monthly
[d] More than 3160 USD monthly
Note: Mother's education is the only control that shows statistically difference between control and treatment groups.

of fit test, where the “theoretical” proportions correspond to the whole sample of the university (7,450) and the realized frequencies corresponded to the reference group (aggregation of the treatment and control group).

The results reported in Table 3 indicate that age is statistically different at a 0.05 significance level, but marginally. However, there is no a substantial difference regarding students' age. In the case of the proportion of women in the sample, we would not reject the null hypothesis of equal proportions. Nonetheless, it is clear that the average income of the families for our sample is higher on average that the one from the university. Finally, we cannot reject that the level of education of the students' mothers from our experiment is different from the ones at university.

In the experiment, the representative student is a 21 years old woman, catholic, with a household monthly income between 1,580 and 3,160 USD, a mother with 16 years of education, does not have a scholarship, has a Scaled C-index equal to 76.75, has scores for average intrinsic motivation of 22, studies for exams 5.5 hours on average per week, is a nonsmoker and consumes alcohol at least once time per week.

Table 3—: Tests for representativeness of the sample

Variable	Reference	University	p-value
Age	20.677	20.153	0.046*
Gender: Male	0.483	0.521	
Gender: Female	0.516	0.478	0.674
Socioeconomic strata: 1	0.032	0.067	
Socioeconomic strata: 2	0.000	0.163	
Socioeconomic strata: 3	0.096	0.222	
Socioeconomic strata: 4	0.096	0.150	
Socioeconomic strata: 5	0.354	0.204	
Socioeconomic strata: 6	0.419	0.191	0.001*
Mother's education: Primary	0.064	0.120	
Mother's education: High School	0.129	0.306	
Mother's education: Technical	0.258	0.139	
Mother's education: Undergraduate	0.419	0.309	
Mother's education: Posgraduate	0.129	0.123	0.079

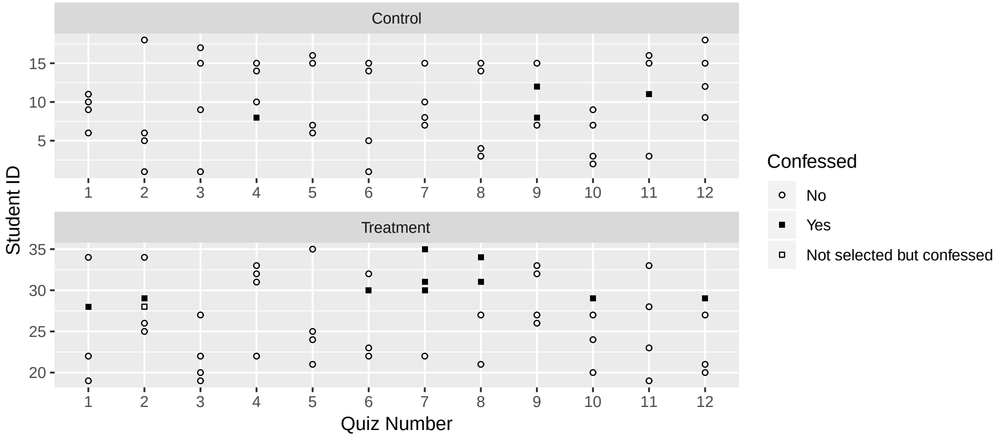
Strata is a Colombia categorization of households to implement social programs. Stratum 1 is the lowest income households, whereas stratum 6 is the highest.

* Difference is statistically significant at 5%

Note: There is statistically significant difference at 1% level in socioeconomic strata between university population and experimental sample.

The disclosure summary is presented in Figure 1. In the control group there were four disclosures, corresponding to three students. In the treatment group, there were eleven disclosures, corresponding to six students. As depicted by Figure 1, most of the disclosures are concentrated in the period after the second half of the experiment.

Figure 1. : Disclosure summary



A. Econometric results

Marginal Analysis. — Our results suggest that attending one additional integrity intervention implies an increase in the probability of disclosing of approximately 4 percentage points. Table 4 depicts summary statistics for this marginal effect. We find that is robust, and statistically significant under different sets of controls. Particularly, once one has controlled for educational controls, the marginal effect is very stable. For the full model, the marginal effect is 3.95, with a 95% credible interval from 2.84 and 4.87, using all the sets of controls. The posterior distribution associated with this marginal effect is depicted in Figure 2. In Table 8 we show that, even though there is some variation, the marginal effect associated with one additional integrity intervention is robust regardless of the model.[11] Table 9 and Figure 6 present convergence diagnostics for the chains regarding each marginal effect in Table 4. The stationarity test proposed by Heidelberger and Welch (1983) and the mean difference test from Geweke et al. (1991) suggest that the chains converge to stationary distributions. Such results are reinforced by the trace plots in Figure 6, which suggest that the chains come from a stationary distribution.

[11] The main difference from the Bayesian longitudinal linear model and the other ones, is that it is the only model controlling for unobserved heterogeneity. In every specification we have used the full set

Table 4—: Marginal effect of number of attended interventions

		95% HDI[+]					
Mean	Median	2.5%	97.5%	*Cog*[a]	*Educ*[b]	*SE*[c]	*H*[d]
2.73	2.71	1.82	3.73	✓			
3.95	3.95	2.98	5.02	✓	✓		
3.92	3.92	2.87	4.95	✓	✓	✓	
3.95	3.95	2.84	4.87	✓	✓	✓	✓

[+] HDI: Highest density interval.
[a] Cognitive controls: C-index and intrinsic motivation .
[b] Educational controls: Mother's education, scholarship (yes or no), belonging to college groups, and study hours for exams.
[c] Socioeconomic controls: Age, female, catholic, and monthly income
[d] Habit controls: smoker, and alcohol drinker.
Note: It seems that the marginal effect associated with the number of interventions is relatively robust to control inclusion.

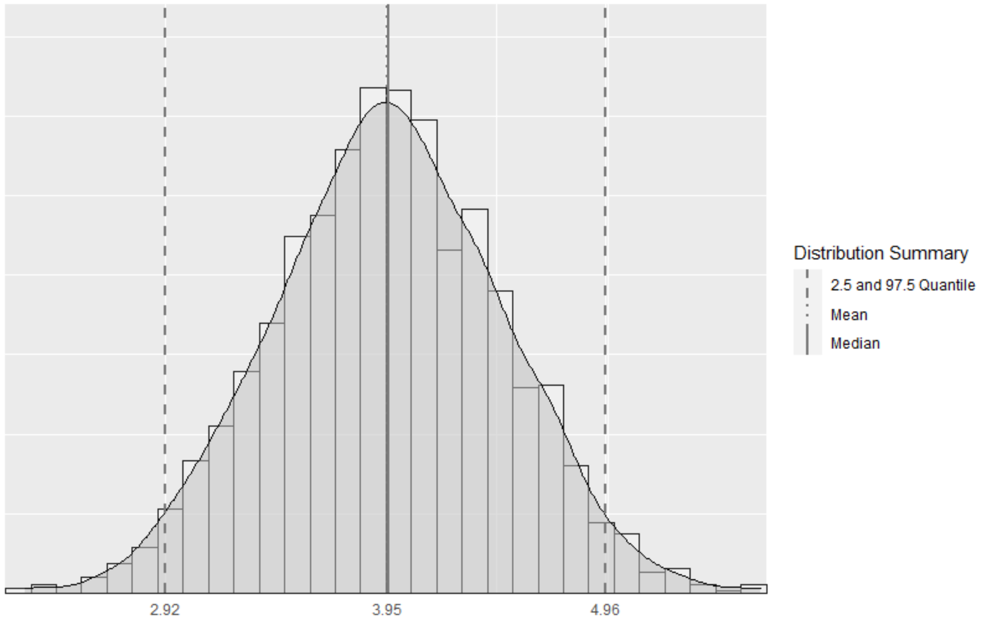
Table 5 depicts the marginal effects associated with the full set of covariates. From now on, significance statements will be performed using a 95% credibility level. We find that both psychological variables have a positive and significant effect on the disclosing. Regarding the C–Index, our results suggest that students who let their judgment of behavior be determined by moral concerns or principles, rather than by other psychological forces, are more prone to disclose the lecturer's mistake. Additionally, we find that students with higher levels of intrinsic motivation, are more likely to disclose. That is, students who value more knowledge than grades, who feel motivated when an exercise or exam challenges them, among others, are more likely to confess. In fact, one of the questions regarding intrinsic motivation was in which degree the students agreed that it is more important to know what were their mistakes in an exam rather than the grade itself. Which makes complete sense with findings from Table 5.

Regarding educational controls we find that neither belonging to college groups nor the weekly hours studied for the subject have a significant effect. On the other hand, we find that students whose mother has a higher level of education are more prone to disclose. We find that the covariate with the biggest magnitude corresponds to owning a scholarship. This could be due to the fact that, students are awarded a scholarship not only for having good grades but for being an example for their peers.

Concerning socioeconomic controls we find that age has no significant effect. In fact, there is little variability regarding this variable. Since it is hard to expect that 1 or 2 years have a substantial effect on students moral behavior, this finding is not surprising. We also find that none of the income dummies are significant. In fact, from our point of view, there is no an intuitive reason to assume that

of controls.

Figure 2. : Posterior distribution of the marginal effect associated with number of attended interventions controlling for all regressors



students with a higher income should be more (or less) honest. We find that females tend to be more honest. In fact, females are around 7% points more likely to disclose than man. Finally, we find that the likelihood of disclosing increases substantially if the student is catholic.

We find that being a cigarette smoker does not affect the disclosing probability. Nonetheless, alcohol consumers are 9% points less likely to disclose. This is an interesting results. We hypothesize that this could be due to the fact that students that do not consume alcohol at least once time a week, are more prone to follow rigorous moral standards, which could make them more prone to disclose the lecturer's mistake. Finally, we find relevant to remark that as Table 5 shows, the degree of variability due to unobserved heterogeneity on average is $\frac{\sigma_b^2}{\sigma_b^2+\sigma^2} = 24\%$, that is, around a quarter of the total variation is associated with unobserved heterogeneity. Since this value is not negligible, this is an argument for taking this decomposition into account in our econometric specification. In other words, it motivates the incorporation of the random effects model instead of estimating a normal standard model.

Given our experimental design, we are not able to disentangle the impact of the interventions separately. Moreover, interventions are not identical and it is difficult to identify which one is effective. Also, there can be lags regarding

Table 5—: Marginal effects for full set of controls

Variable	Mean	Median	95% CI	
Intercept	-175.15	-175.25	-244.22	-105.98
Number of Interventions	3.95	3.95	2.84	4.87
C-index	0.59	0.59	0.37	0.81
Intrinsic Motivation	1.55	1.55	0.45	2.66
Mother's education	3.31	3.30	2.20	4.54
Scholarship	20.03	20.07	5.85	32.58
College groups	-1.41	-1.39	-10.80	6.51
Study Hours	1.58	1.60	-0.23	3.42
Age	1.27	1.31	-1.09	4.14
Female	6.81	6.89	1.38	12.44
Catholic	16.07	15.97	2.59	30.60
Income 1	4.78	4.39	-13.48	21.52
Income 2	-14.72	-14.92	-35.37	8.64
Income 3	15.11	15.16	-2.18	34.17
Smoker	9.22	9.28	-9.66	26.51
Alcohol Consumer	-9.08	-9.07	-15.92	-2.86
σ_b^2	3.22	2.95	0.98	5.83
σ^2	7.78	7.65	5.09	10.51
$\frac{\sigma_b^2}{\sigma_b^2+\sigma^2}$	23.76%	23.46%	8.51%	39.95%

Note: Most of the controls are statistically significant at 95% level. Having a scholarship is the most relevant control.

effects, some may be more effective than others, and the order may matter. We pretend to evaluate the campaign as a whole, not particular interventions. Just taking into account the order, there will be 12! ≈ 479 million ways. However, we perform two different exercises. In the first one, we estimate the full model (row 4 in Table 4) including the square of attended interventions and a time trend. The first variable allows us to capture non-linearity in the effect and potentially finding an optimum. The second one is a proxy for potential spillover effects. The second exercise is to estimate the full model, including a dummy variable (1 if the student belongs to the treatment group and 0 if it belongs to the control) instead of the number of attended interventions. This would give us an idea of the effect of the campaign as a whole. Results for these exercises are reported in Table

6. From the first exercise, we conclude that the effect of the interventions is not linear. Moreover, there is a maximum at 6.4 interventions, which would suggest that between 6 and 7 interventions could be enough to elicit honest behavior among students in our campaign. From the second exercise, we find that, on average, the integrity campaign increases the probability of disclosing by around 32 percentage points. Moreover, we find that the time trend (which we use as a proxy for potential spillovers) is not significant at a 0.05 significance level in none of the models. Regarding potential spillover effects, what could have happened is that students from the treatment group had a positive effect on the control group, making them more prone to disclose. So at most, if there is a spillover effect, it would have caused an attenuation bias.

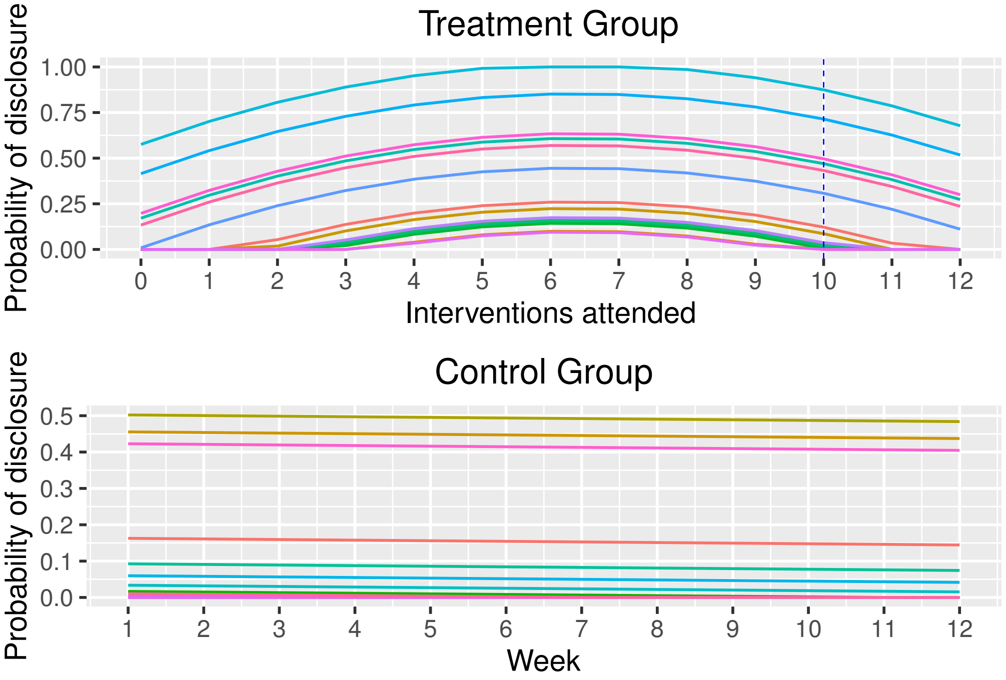
Table 6—: Robustness check for the marginal effect of the integrity campaign

	(1)	(2)
Number of attended interventions	13.680	
	(11.265, 16.235)	
Number of attended interventions squared	-1.069	
	(-1.352, -0.782)	
Time trend	-0.166	0.462
	(-0.750, 0.482)	(-0.078, 1.024)
Treatment		31.911
		(25.951, 38.035)

Note: Point estimates correspond to posterior means, and values in brackets correspond to the highest density posterior Intervals with a 95% credible mass.
It seems that there are not spillover effects as the time trend is not statistically significant, and the treatment group increases the probability of disclosure in 32% compared to the control group.

Finally, we study how the likelihood of disclosures changes with time. For example, at the end of the experiment, students in the treatment group might be more likely to disclose, due to the fact that they have attended more interventions. On the other hand, if there were spillover effects, students in the control group might be more likely to disclose. To study this issue we calculate the probability of disclosure for every student using the estimates from model 1 in Table 6. Since none of the student in the control group attended to any interventions, we calculate the probability of disclosure as a function of the time trend. For the treatment group, we calculate the probability of disclosure as a function of attended interventions. We can see in Figure 3 that the control group has a probability of disclosing virtually constant. If there were spillover effects, the probability of disclosing should be upward sloping. For the treatment group, we find that the highest probabilities are around 6 and 7 interventions, as it is suggested by the previous results. There is a vertical line at 10 interventions since this is the highest value we have in the sample. Notice that, even though the probability of disclosing decreases after the 6-7 intervention, the value it attains at 10 interventions is higher for every student in the treatment group. Such result indicates a positive effect of the integrity interventions on the probability of disclosure.

Figure 3. : Probability of disclosing for every student



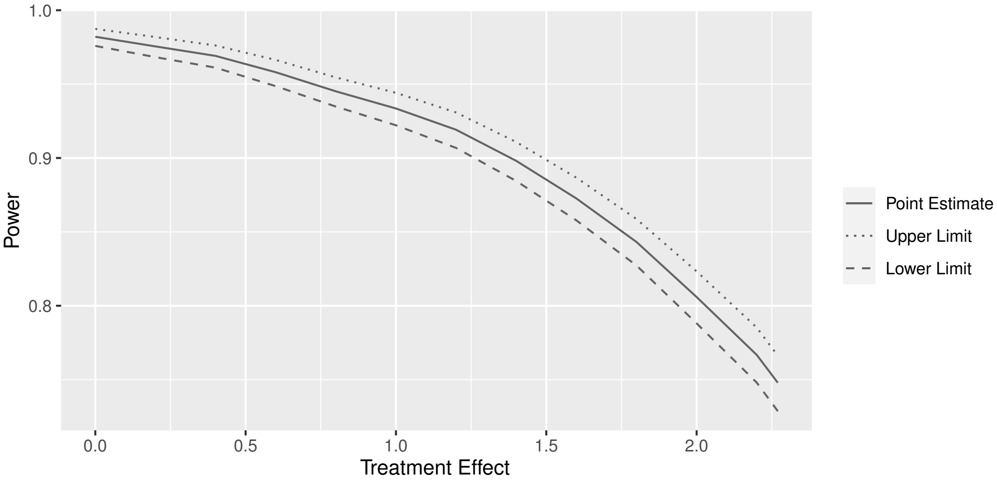
Retrospective Power Analysis. — In the light of the small sample size, we perform Algorithm A1 to estimate the probability that the interventions have an effect on the probability of disclosing greater than a range of values. We can observe the power analysis between 0 (H_0. No marginal effect) to 2.61 (H_0. marginal effect=2.61) in Figure 4 and Table 7. The analysis shows that the probability of rejecting the null hypothesis of no effect given that this is false is 98.2% (there is 95% probability that the value is in the interval (97.5%, 98.7%)). The power of the null hypothesis of a marginal effect equal to 1 point is 93.3% (there is 95% probability that the value is in the interval (92.2%, 94.4%)), and marginal effect equal to 2.27 points is 74.8% (there is 95% probability that the value is in the interval (72.9%, 76.6%)). In general, we can see that the power is high for the null hypothesis of no effect given that this is false, and it decreases as we approach the minimum marginal effect estimate.

Table 7—: Numerical results: Bayesian power analysis

	0	0.2	0.4	0.6	0.8	1	1.2	1.4	1.6	1.8	2	2.2	2.27
Point estimate	0.98	0.98	0.97	0.96	0.95	0.93	0.92	0.90	0.87	0.84	0.81	0.77	0.75
Lower limit	0.98	0.97	0.96	0.95	0.93	0.92	0.91	0.88	0.86	0.83	0.79	0.75	0.73
Upper limit	0.99	0.98	0.98	0.97	0.95	0.94	0.93	0.91	0.89	0.86	0.82	0.79	0.77

Note: Mean, 2.5% and 97.5% percentiles of the power analysis. It seems that power is good despite the small sample size.

Figure 4. : Bayesian Power test



B. Qualitative results

Undergraduate econometrics students participated in a focus group. The results mentioned above are in line with what they said in the post-experimental feedback. Among the things they mentioned, one of the possible motives for not disclosing is the lack of interest in really learning in college. Another reason they gave for not disclosing was that if they really needed the grade, they would not reject it; actually, one of the questions regarding intrinsic motivation was if they valued more the acquisition of knowledge than the grade itself. According to our findings, getting less points concerning intrinsic motivation decreased the probability of disclosing.

Other important factors which students remarked on regarding the decision to disclose were: the behavior of their peers (specially their close ones), the use of self-regulatory mechanisms in order to avoid acting dishonestly, and the attitude of the lecturer toward them and the class.

V. Conclusions

Economics teaching should not be based just on technical aspects. Integrity is a fundamental component in human development, and should be reinforced in college education. This is particularly important in economics due to the important role of economists in public and private positions. Our experiment suggests that, at least in the short term, interventions that promote integrity can have positive moderate effects on students' behavior regarding this aspect. Our Bayesian power analysis shows that this result seems to be robust despite the small sample size.

This positive outcome was achieved in an environment where the external reward was very high. In particular, the students mentioned, in the post-experimental focus group, the difficulty of the subject. So, it seems that the awareness of moral standards pushes internal rewards up, as a consequence, promotes integrity. This is a hopeful outcome in a society having cultural factors that promote being sly.

Therefore, our results suggest that an integrity campaign based on the kind of interventions that we set should be done as this seems to have positive effects on undergraduate students' academic integrity.

Future research should consider the role of incentives on cheating. This is a drawback of our setting as there is no variation of benefits for participants. This variation could be exploited to investigate how different incentives may modify students' behaviour.

References

Anderman, Eric M., and Carol Midgley. 2004. “Changes in self-reported academic cheating across the transition from middle school to high school.” *Contemporary Educational Psychology*, 29(4): 499–517.

Anderman, Eric M., Tripp Griesinger, and Gloria Westerfield. 1998. “Motivation and cheating during early adolescence.” *Journal of Educational Psychology*, 90(1): 84.

Ayal, Shahar, Francesca Gino, Rachel Barkan, and Dan Ariely. 2015. “Three principles to REVISE people’s unethical behavior.” *Perspectives on Psychological Science*, 10(6): 738–741.

Bandura, Albert. 2002. “Selective moral disengagement in the exercise of moral agency.” *Journal of Moral Education*, 31(2): 101–119.

Bandura, Albert, Claudio Barbaranelli, Gian Vittorio Caprara, and Concetta Pastorelli. 1996. “Mechanisms of moral disengagement in the exercise of moral agency.” *Journal of Personality and Social Psychology*, 71(2): 364.

Barkan, Rachel, Shahar Ayal, and Dan Ariely. 2015. “Ethical dissonance, justifications, and moral behavior.” *Current Opinion in Psychology*, 6: 157–161.

Bonetti, Shane. 1998. “Experimental economics and deception.” *Journal of Economic Psychology*, 19(3): 377–395.

Bowers, William J. 1964. *Student dishonesty and its control in college.* . 1 ed., New York:Stanford University Libraries.

Cohen, Jacob. 2013. *Statistical power analysis for the behavioral sciences.* Routledge.

Davis, Stephen F, Cathy A Grover, Angela H Becker, and Loretta N McGregor. 1992. “Academic dishonesty: Prevalence, determinants, techniques, and punishments.” *Teaching of Psychology*, 19(1): 16–20.

Finn, Kristin Voelkl, and Michael R. Frone. 2004. “Academic performance and cheating: Moderating role of school identification and self-efficacy.” *Journal of Educational Research*, 97(3): 115–121.

García-Villegas, Mauricio, Nathalia Franco-Pérez, and Alejandro Cortés. 2016. “Perspectives on academic integrity in Colombia and Latin America.” In *Handbook of Academic Integrity.* , ed. Tracey Bretag, 161–180. Singapore:Springer.

G. Charness, A. Samek, and J. van De Ven. 2021. “What Is Deception in Experimental Economics? A Survey.”

https://benny.aeaweb.org/conference/2021/preliminary/1934?q=eNqrVipOLS7OzM8LqSxIVbKqhnGVrJQMlWp1lBKLi_OTgRwlHaWS1KJcXAgrJbESKpSZmwphlWWmloOOFxUUXDAFTA1AegsSOOGyhkAOXDCMpB5E, [Online; accessed 30-August-2021].

Gerard, Patrick D, David R Smith, and Govinda Weerakkody. 1998. "Limits of retrospective power analysis." *The Journal of wildlife management*, 801–807.

Geweke, John, et al. 1991. *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments.* Vol. 196, Federal Reserve Bank of Minneapolis, Research Department Minneapolis, MN.

Gneezy, Uri. 2005. "Deception: The role of consequences." *American Economic Review*, 95(1): 384–394.

Greenberg, Edward. 2008. *Introduction to Bayesian Econometrics.* Cambridge University Press.

Grimes, Paul W. 2004. "Dishonesty in academics and business: A Cross-cultural evaluation of student attitudes." *Journal of Business Ethics*, 49(3): 273–290.

Grimes, Paul W., and Jon P. Rezek. 2005. "The determinants of cheating by high school economics students: A comparative study of academic dishonesty in the transitional economies." *International Review of Economics Education*, 4(2): 23–45.

Haines, Valerie J., George M. Diekhoff, Emily E. LaBeff, and Robert E. Clark. 1986. "College cheating: Immaturity, lack of commitment, and the neutralizing attitude." *Research in Higher Education*, 25(4): 342–354.

Heidelberger, Philip, and Peter D Welch. 1983. "Simulation run length control in the presence of an initial transient." *Operations Research*, 31(6): 1109–1144.

Hoenig, John M, and Dennis M Heisey. 2001. "The abuse of power: the pervasive fallacy of power calculations for data analysis." *The American Statistician*, 55(1): 19–24.

Hyman, Ira E., S. Matthew Boss, Breanne M. Wise, Kira E. McKenzie, and Jenna M. Caggiano. 2010. "Did you see the unicycling clown? Inattentional blindness while walking and talking on a cell phone." *Applied Cognitive Psychology*, 24(5): 597–607.

Index, Corruption Perceptions. 2019. "Corruption perceptions index 2018." *Transparancy International.*

Jordan, Jennifer, Elizabeth Mullen, and J Keith Murnighan. 2011. "Striving for the moral self: The effects of recalling past moral actions on future moral behavior." *Personality and Social Psychology Bulletin*, 1–13.

Kanat-Maymon, Yaniv, Michal Benjamin, Aviva Stavsky, Anat Shoshani, and Guy Roth. 2015. "The role of basic need fulfillment in academic dishonesty: A self-determination theory perspective." *Contemporary Educational Psychology*, 43: 1–9.

Kelman, Herbert C. 1967. "Human use of human subjects: The problem of deception in social psychological experiments." *Psychological Bulletin*, 67(1): 1.

Khodaie, Ebrahim, Ali Moghadamzadeh, and Keyvan Salehi. 2011. "Factors affecting the probability of academic cheating school students in Tehran." *Procedia-Social and Behavioral Sciences*, 29: 1587–1595.

Kohlberg, Lawrence, and Richard H Hersh. 1977. "Moral development: A review of the theory." *Theory Into Practice*, 16(2): 53–59.

Kruschke, John. 2014. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan.* Academic Press.

Kruschke, John K. 2013. "Bayesian estimation supersedes the *t* test." *Journal of Experimental Psychology: General*, 142(2): 573.

Kruschke, John K., and Torrin M. Liddell. 2018. "The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective." *Psychonomic Bulletin & Review*, 25(1): 178–206.

Lind, G. 2008. "The meaning and measurement of moral judgment competence." In *Contemporary Philosophical and Psychological Perspectives on Moral Development and Education.* . 3 ed., , ed. Jr. Daniel Fasko and Wayne Willis, Chapter 8, 185–220. Creskill: Hampton Press. An optional note.

Lind, Georg. 2007. "Scoring of the Moral Judgment Test (MJT)." *Journal of Military Ethics Special Issue: Moral Judgement Within the Armed Forces*, 6(1): 19–40.

Lipson, Angela, and Norma McGavern. 1993. "Undergraduate academic dishonesty at MIT: Results of a study of attitudes and behaviour of undergraduates." The institution that published, Massachusetts Institute of Technology. https://files.eric.ed.gov/fulltext/ED368272.pdf.

López-Pérez, Raúl, and Eli Spiegelman. 2013. "Why do people tell the truth? Experimental evidence for pure lie aversion." *Experimental Economics*, 16(3): 233–247.

Martin, Andrew D., Kevin M. Quinn, and Jong Hee Park. 2011. "MCMCpack: Markov Chain Monte Carlo in R." *Journal of Statistical Software*, 42(9): 22.

Martinelli, César, Susan W Parker, Ana Cristina Pérez-Gea, and Rodimiro Rodrigo. 2018. "Cheating and incentives: Learning from a policy experiment." *American Economic Journal: Economic Policy*, 10(1): 298–325.

Mazar, Nina, On Amir, and Dan Ariely. 2008. "The dishonesty of honest people: A theory of self-concept maintenance." *Journal of Marketing Research*, 45(6): 633–644.

McCabe, Donald L, and Linda Klebe Trevino. 1997. "Individual and contextual influences on academic dishonesty: A multicampus investigation." *Research in Higher Education*, 38(3): 379–396.

McCabe, Donald L., Kenneth D. Butterfield, and Linda Klebe Trevino. 2006. "Academic dishonesty in graduate business programs: Prevalence, causes, and proposed action." *Academy of Management Learning & Education*, 5(3): 294–305.

McCabe, Donald L, Linda Klebe Treviño, and Kenneth D Butterfield. 2001. "Cheating in academic institutions: A decade of research." *Ethics &Behavior*, 11(3): 219–232.

Nakagawa, Shinichi, and T Mary Foster. 2004. "The case against retrospective statistical power analyses with an introduction to power analysis." *Acta ethologica*, 7(2): 103–108.

O'Keefe, Daniel J. 2007. "Brief report: post hoc power, observed power, a priori power, retrospective power, prospective power, achieved power: sorting out appropriate uses of statistical power analyses." *Communication methods and measures*, 1(4): 291–299.

Ortmann, Andreas, and Ralph Hertwig. 2002. "The costs of deception: Evidence from psychology." *Experimental Economics*, 5(2): 111–131.

Park, Chris. 2003. "In other (people's) words: Plagiarism by university students–literature and lessons." *Assessment & evaluation in higher education*, 28(5): 471–488.

Piaget, Jean. 2013. *Child's Conception of Number: Selected Works.* Vol. 2, Routledge.

Pontificio, Anuario. 2014. "Annuarium Statisticum Ecclesiae 2012." *L'Osservatore romano*, 8(8): 2014.

Ryan, Richard M., and Edward L. Deci. 2000. "Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being." *American Psychologist*, 55(1): 68.

Shaughnessy, John J., and Eugene B. Zechmeister. 2012. *Research Methods in Psychology.* Knopf.

Spiegelhalter, D., T. Best, W. Gilks, and D. Lunn. 2003. "BUGS: Bayesian inference using Gibbs sampling." MRC Biostatistics Unit, England. www.mrc-bsu.cam.ac.uk/bugs/.

Steidl, Robert J, John P Hayes, and Eric Schauber. 1997. "Statistical power analysis in wildlife research." *The Journal of Wildlife Management*, 270–279.

Sun, Shuyan, Wei Pan, and Lihshing Leigh Wang. 2011. "Rethinking observed power." *Methodology.*

Thaler, Richard. 1980. "Toward a positive theory of consumer choice." *Journal of Economic Behavior & Organization*, 1(1): 39–60.

Thomas, Len. 1997. "Retrospective power analysis." *Conservation Biology*, 11(1): 276–280.

Valentino, Kristin, Dante Cicchetti, Sheree L Toth, and Fred A Rogosch. 2011. "Mother–child play and maltreatment: A longitudinal analysis of emerging social behavior from infancy to toddlerhood." *Developmental Psychology*, 47(5): 1280.

Wooldridge, Jeffrey M. 2010. *Econometric analysis of cross section and panel data.* MIT Press.

VI. Appendix

A. Appendix A: Intervention details

We give some details about interventions shown in Table 1:

- Week 1: The center of integrity of the university gave a 15 minutes talk about ethics and academic integrity explaining these concepts and giving some examples and potential consequences. In addition, the quiz in that week had a heading with the 10 commandments for the treatment group. The control group quiz did not have that heading.

- Week 2: We played a fragment of the documentary "Inside Job" where it is explained the origin and consequences of the financial crisis in 2008. This documentary shows how integrity lack was basically the origin of this situation.

- Week 3: The integrity center of the university gave a 15 minutes talk about integrity in the professional environment. What professional integrity is, civil and legal consequences of lack of professional integrity, and some relevant and well-known local examples.

- Week 4: The previous weekly quiz is solved in the next applied class after delivering the marked quiz. However, we asked students in the treatment group to come over to lecture's office this week, get their quiz, and ask them directly if the grade was OK, and talk about their satisfaction in class.

- Week 5: We asked the treatment group to sign a "honesty code" that says: "As student of this university, I promise to do all academic activities with honesty, that is, without any cheating, and citing all references. Moreover, I certify that integrity, responsibility and respect will guide my way through academic duties such as exams, quizzes, writing tasks and presentations as well as any other evaluating assignment in my program".

  In addition, the center of integrity of the university gave a 15 minutes talk about an institutional integrity campaign called "Atreverse a Pensar" that was implemented during 2011 (this experiment was conducted in 2016).

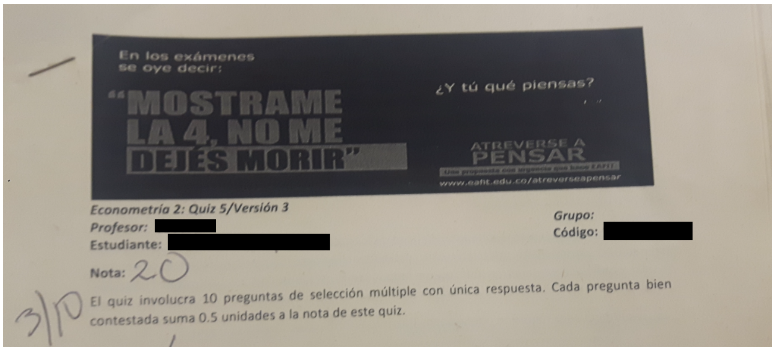
- Week 6: The integrity center gave to each student in the treatment group a bracelet saying "Dare to think". This was the slogan of the integrity campaign in 2011. We also named the student who disclosed lecture's marking mistake, and thank her/him for this act.

- Week 7: We played the fist part of the episode "Trolley Problem and Kin Selection" of the series "Brain Games" where some moral dilemmas are shown.

- Week 8: We asked treatment group students to come over to lectures' office to get their quizzes, and talk about students' course performance.
- Week 9: We discussed in class a local very well-known case of corporate dishonesty. Students gave their opinions about the subject.
- Week 10: We played the second part of the episode "Trolley Problem and Kin Selection" of the series "Brain Games" where some moral dilemmas are shown.
- Week 11: The integrity center asked a former student from the university, who was expelled out due to supplanting another student in a final exam, to give testimony about his behaviour.
- Week 12: Students in the treatment group talked about dishonest acts in her/his lives: why they did this, how they feel about that, etc.

B. *Appendix B: Figures and Tables*

Figure 5. : Quiz grading scheme.



En los exámenes
se oye decir:

"MOSTRAME
LA 4, NO ME
DEJÉS MORIR"

¿Y tú qué piensas?

ATREVERSE A
PENSAR

Una propuesta con urgencia que hace EAFIT

www.eafit.edu.co/atreverseapensar

Econometría 2: Quiz 5/Versión 3

Profesor: [redacted]

Estudiante: [redacted]

Grupo:

Código: [redacted]

Nota: 20

3/10

El quiz involucra 10 preguntas de selección múltiple con única respuesta. Cada pregunta bien contestada suma 0.5 unidades a la nota de este quiz.

C. *Appendix C: Posterior distributions*

Following the Bayes' rule:

$$\pi(\beta, b_i, \sigma^2, \sigma_b^2 | y_{it}) \propto f(y_{it}|\beta, b_i, \sigma^2, \sigma_b^2)\pi(\beta)\pi(\sigma^2)\pi(\sigma_b^2).$$

Figure 6. : Trace plots for marginal effects for each model in Table 4

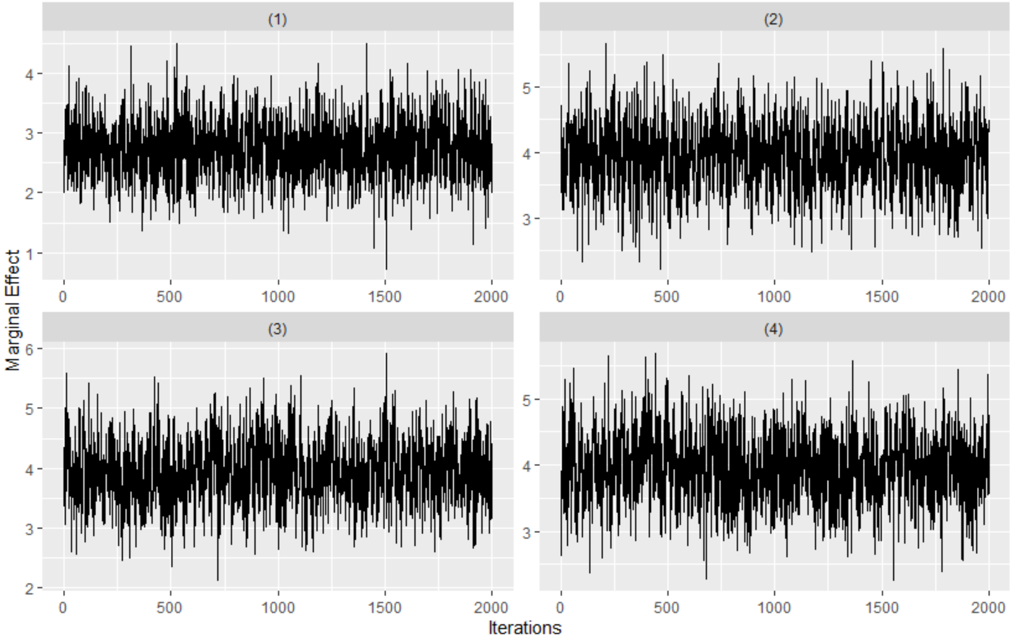


Table 8—: Marginal effect of number of attended interventions: Robustness checks

Model	Point estimate	95% CI	
Bayesian longitudinal linear	3.95	2.84	4.87
Frequentist probit	4.32	1.25	7.39
Bayesian linear	4.92	1.96	7.70
Frequentist linear	4.92	2.09	7.75

Taking into account that $\boldsymbol{y}_i|\boldsymbol{\beta},\sigma_b^2,\sigma^2 \sim \mathcal{N}(\boldsymbol{X}_i\boldsymbol{\beta},\boldsymbol{V}_i)$ where $\boldsymbol{V_i} = \sigma^2\boldsymbol{I}_{n_i} + \sigma_b^2\boldsymbol{i}_{n_i}\boldsymbol{i}'_{n_i}$ and $\boldsymbol{i}_{n_i}$ is a vector of 1's, then

$$y_{ij}|\boldsymbol{\beta},\sigma^2,\boldsymbol{b},\boldsymbol{y},\boldsymbol{X},\boldsymbol{W} \propto \mathcal{N}(\boldsymbol{x}'_{ij}\boldsymbol{\beta}+b_i,\sigma^2),$$

$$\boldsymbol{\beta}|\sigma^2,\sigma_b^2,\boldsymbol{y},\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\beta}^*,\boldsymbol{B}),$$
$$b_i|\boldsymbol{\beta},\sigma^2,\sigma_b^2,\boldsymbol{y},\boldsymbol{X} \sim \mathcal{N}(b_i^*,s_i^*),$$

Table 9—: Convergence diagnosis for marginal effects of Table 4

Geweke[b]	Heidelberg[a]
1.46	0.64
-0.44	0.97
-0.29	0.19
0.76	0.12

[a] Null hypothesis is stationarity of the chain (p-values)

[b] Mean difference test z-score Dependence factor

$$\sigma_b^2|\boldsymbol{b} \sim \mathcal{IG}(d^*, v^*),$$

$$\sigma^2|\boldsymbol{\beta}, \sigma_b^2, \boldsymbol{b}, \boldsymbol{y}, \boldsymbol{X} \sim \mathcal{IG}(\alpha^*, \delta^*),$$

where $\boldsymbol{B} = (\boldsymbol{B}_0^{-1} + \sigma^{-2}\sum_{i=1}^n \boldsymbol{X}_i'\boldsymbol{V}_i^{-1}\boldsymbol{X}_i)^{-1}$, $\boldsymbol{\beta}^* = \boldsymbol{B}(\boldsymbol{B}_0^{-1}\boldsymbol{\beta}_0 + \sigma^{-2}\sum_{i=1}^n \boldsymbol{X}_i'\boldsymbol{V}_i^{-1}\boldsymbol{y}_i)$, $s_i^* = (\sigma_b^{-2} + n_i\sigma^{-2})^{-1}$, $b_i^* = s_i^*(\sigma^{-2}\boldsymbol{i}_{n_i}'(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}))$, $d^* = \frac{d_0+m}{2}$ and $v^* = 2d_0v_0 + 2\sum_{i=1}^m b_i^2$, $\alpha^* = \alpha_0 + \frac{1}{2}\sum_{i=1}^m n_i$ and $\delta^* = 1/\delta_0 + \frac{1}{2}\sum_{i=1}^m(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{i}_{n_i}b_i)'(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{i}_{n_i}b_i)$.

An advantage of this hierarchical longitudinal model is that it does not require either the panel to be balanced or equidistant, it just requires that several units are observed at different points in time.

D. *Appendix D: Retrospective power analysis*

Algorithm A1 Bayesian power analysis

```
Estimate a hierarchical random effects normal model according to Equation (1), and obtain the chains from the posterior distributions.
Get $s = 1, \dots, S$ parameters from the posteriors, $S$ is the effective dimension of the posterior chains (discarding the burn-in and taking into account the thin parameter).
Define $\beta_{int}^{(s)}$ as the marginal effect associated with the number of attended interventions at the s-th step
Define a series of thresholds $\beta_{int}^{0,m}$ with $m = 1, \dots, M$ where $\beta_{int}^{0,M} = min\{\beta_{int}^{(1)}, \dots, \beta_{int}^{(S)}\}$. We set the thresholds as a sequence of values from 0 to $\beta_{int}^{0,M}$ with a step of 0.2
Simulate the dependent variable using the observed data following the same structure, that is, taking into account $m$ and $n_i$, as follows:
for $s = 1, \dots, S$ do
    • Draw the unobserved heterogeneity $b_i^{(s)}$ from a normal distribution with mean 0 and variance $\sigma_b^{2(s)}$.
    • Draw the stochastic error $\epsilon_{it}^{(s)}$ from a normal distribution with mean 0 and variance $\sigma^{2(s)}$.
    • Compute $y_{it}^{(s)}$ as $y_{it}^{(s)} = \boldsymbol{x}_{it}'\boldsymbol{\beta}^{(s)} + b_i^{(s)} + \epsilon_{it}^{(s)}$.
    • Estimate a Bayesian longitudinal random effects normal model using $y_{it}^{(s)}$ and $\boldsymbol{x}_{it}$, whose parameters in the posterior chains are indexed by $l$, with $l = 1, \dots, L$.
for $s = 1, \dots, S$ do
    for $m = 1, \dots, M$ do
        Define $\beta_{int}^{s,5}$ as the 5-th percentile of the highest posterior density interval of $\{\beta_{int}^{(s,1)}, \dots, \beta_{int}^{(s,L)}\}$
        Get $P$ realizations of the posterior distribution for $P_{H_a}^s(\beta_{int}^{0,m})$ using a Beta-Binomial model with a noninformative beta prior ($P_{H_a}^s(\beta_{int}^{0,m}) \sim \mathcal{B}eta(1,1)$). This implies $P_{H_a}^s(\beta_{int}^{0,m})|\boldsymbol{y}, \boldsymbol{X} \sim \mathcal{B}eta\,(\alpha, \beta)$ Where $\alpha = 1 + \sum_{s=1}^{S} 1\left[\beta_{int}^{s,5} \geq \beta_{int}^{0,m}\right]$ and $\beta = 1 + S - \sum_{s=1}^{S} 1\left[\beta_{int}^{s,5} \geq \beta_{int}^{0,m}\right]$.
        This corresponds to the probability of rejecting the null hypothesis given that it is false. It is false because $P(\beta_{int}^{0,m} < min\{\beta_{int}^{(1)}, \dots, \beta_{int}^{(S)}\}) = 1$ under the posterior distribution by construction $\forall s$ and $\forall m$.
        Obtain the point estimate of the power $P_{H_a}\left(\beta_{int}^{0,m}\right)$ as the mean of the the posterior of $P_{H_a}\left(\beta_{int}^{0,m}\right)$.
        Obtain the credible interval for the power as the 2.5th and 97.5th percentiles of the highest posterior density interval corresponding to the posterior of $P_{H_a}\left(\beta_{int}^{0,m}\right)$
```