ERASMUS UNIVERSITY ROTTERDAM
Erasmus School of Economics
Master Thesis [Economics & Business – Behavioural Economics]

Betting on others to reveal oneself

An experimental test of bets to elicit private signals in student evaluations of teaching

Name student: Jonas Cornelius Fuerle
Student ID number: 484500

Supervisor: Yan Xu
Second assessor: Aurelien Baillon

Date final version: 15.08.2019

The views stated in this thesis are those of the author and not necessarily those of Erasmus School of Economics or Erasmus University Rotterdam

Abstract

This thesis tests Top/Flop and Target betting, a new approach to elicit agents' private signals when the truth is unverifiable. The method elicits an agent's private signal based on the relative performance of an item against a comparable but unknown item using bets to incentivise subjects. We test this method in the domain of student evaluations of teaching under the implementation of an exogenous default bias. Our results indicate that Top/Flop and Target betting can increase data quality in regard to subject's response time. Further, upon testing the assumptions of the method, we conclude that they are reasonable to hold in practice. These experimental outcomes underscore the discussion of finding simpler methods in the Bayesian truth-inducing incentive literature. Our results might help in designing more reliable methods of opinion elicitation and encourage further testing of the Top/Flop and Target betting method in large-scale sample populations.

Keywords: Private signals; Student evaluations of teaching; opinion elicitation; Bayesian game; Bets

Table of Contents

ABSTRACT 1

1 INTRODUCTION 3

2 RELATED LITERATURE 4

2.1 RESEARCH QUESTION & HYPOTHESES 8

3 EXPERIMENTAL DESIGN 9

3.1 TOP/FLOP AND TARGET BETTING - DISCUSSING THE ASSUMPTIONS 9
3.2 SUBJECTS AND EXPERIMENTAL PROCEDURE 10
3.3 STAGE 2 - EXPERIMENTAL TREATMENTS 11

4 RESULTS 14

4.1 TOP/FLOP BETTING RESULTS 16
4.1.1 Payment Mechanism & Evaluation Outcomes 16
4.1.2 Default Bias 17
4.1.3 Drop-out Rates 18
4.1.4 Response Time 19
4.1.5 Bet Valuation 21
4.2 TARGET BETTING RESULTS 22
4.2.3 Drop-out Rates 24
4.2.4 Response Time 24
4.2.5 Bet Valuation 26

5 BALANCE & ROBUSTNESS CHECKS 27

5.1 BALANCE CHECK 27
5.2 ROBUSTNESS CHECK 28
5.2.1 Top/Flop Betting 28
5.2.2 Target Betting 30

6 DISCUSSION 31

7 CONCLUSION 32

8 REFERENCES 33

APPENDIX 36

1 Introduction

Course evaluations at the Erasmus School of Economics serve as an important indicator of the state of teaching and student satisfaction. At the end of each Block, students can state their opinion on the courses they have taken. These student evaluations of teaching (SET) cover different aspects ranging from student's subjective evaluation of the course materials to the Professors teaching performance. Moreover, the outcomes of SETs are utilized for decision making purposes, such as improving courses, and to support tenure and promotion decisions for professors. Given the emphasis of these evaluations, turnout historically is underwhelming[1] and reservations with self-selection bias arguably persist (e.g. evaluations as an instrument to vent negative emotions regarding the examination or course (Marlin 1987)).

In general, the elicitation of subjective data relating to an individual's happiness, satisfaction or opinion, is an area of scrutiny among scholars (Bertrand & Mullainathan 2001). Since the researcher cannot credibly determine a truthful response, the subsequent use of subjective data could be undermined. In data elicitation, we can elicit verifiable and unverifiable truths, the former being less problematic in determining validity and the latter being strictly dependent on an individual's truthfulness. Ongoing efforts in the Experimental Economics incentive literature accentuate the prevailing issues with subjective data elicitation. Methods such as the Bayesian-Truth-Serum (Prelec 2004), the Peer-Prediction method (Miller, Resnik & Zeckhauser 2005) and Bayesian markets (Baillon 2017) provide the theoretical basis for the incentive compatible elicitation of subjective truths.

Another new method, Top/Flop- and Target-betting, takes an intuitive approach by using the everyday convention of betting on (currently) unverifiable outcomes (Xu & Baillon 2019). Consider a student who provides her private signal regarding her current course by placing either a Top- or Flop-bet. A Top-bet states that the current course is rated higher than a random other comparable course. Facing the choice between a Top- and Flop-bet, the student will choose the bet which she believes is most likely to win the reward. If she likes the current course (positive signal), she will consider the Top-bet more likely to obtain the reward, whereas if she dislikes the course (negative signal) her subjective probability of the Flop-bet supersedes the Top-bet. This method could overcome the general Keynesian beauty contest type of complications in opinion elicitation, since the subject now compares the relative performance of two similar courses based on her private signal.

Following the motivation to improve student course evaluations, it is logical to wonder whether employing truth-telling incentives in the Erasmus School of Economics SETs yields better quality data. The novel truth-inducing methods aim to improve response rates and data quality as a result of the incentives derived

[1] Regrettably, data regarding course evaluations are protected under the GDPR and were not to be obtained by the Author. However, suggestive evidence is available from two courses due to personal involvement as a Student Assistant. The number of questionnaires that have been completed for each of these two courses ranges between 19 and 40 students where roughly 800 students are eligible to respond.

from Bayesian reasoning. Erasmus School of Economics should be motivated to improve data quality in its student course evaluations because the data is used in their decision-making process. For this reason, this paper investigates whether Top/Flop- and Target-betting incentivization is successful in improving data quality for student course evaluations under the presence of a default bias. We find that TFTB incentivization increases data quality with regards to response time and that subjects choose the bet that maximizes their subjective chance of winning the price. On the contrary, we do not find evidence for a reduction in default bias and drop-out rates.

The rest of the paper is organized as follows. Firstly, we provide a snapshot of the relevant truth-inducing incentive literature, followed by the Top/Flop and Target betting method to elicit unverifiable truths. Next, we present the research question and hypotheses of the experimental setup. Section X covers the experimental design in detail, outlining the subject pool, as well as discussing the assumptions of Top/Flop- and Target-betting methods and the operationalization of the default bias. Afterwards, Section X discusses the results of the experiment by analysing several indicators of data quality such as response time, dropout rates and personal valuation of the betting choice. Finally, we discuss the implications of the experimental outcome for the truth-telling incentive literature and the design of student course evaluations at Erasmus School of Economics and highlight possible future applications of Top/Flop & Target betting incentives.

2 Related Literature

Economics research often relies on the elicitation of subjective beliefs at the individual level (Baillon 2015). Considering the policy implications which might be inferred from respondents stated beliefs, it is paramount to ensure data quality in subjects' responses to questions where the truth is subjective and thus unverifiable by the researcher (Prelec 2004). The discussion of unverifiable truths is important due to the further interconnectedness of Economic Science with other Social Sciences, resulting in an enlarged scope of Economic Science as a whole (Coase 1978; Stigler 1984). These extensions require implementations of question types measuring subjective attitudes, experiences and beliefs, a methodological approach that remains scrutinized in Economics (Bertrand & Mullainathan 2001). Measuring subjective attitudes and beliefs correlates to a central problem in Psychology and Behavioural Economics. Whereas other social sciences employ subjective belief questions without incentivization, economists are keen on following the economic paradigm that incentives matter in increasing individuals' effort, performance and persistence (Camerer & Hogarth 1999).

According to Miller, Resnik and Zeckhauser (2005) subjective truth elicitation methods must overcome two distinct problems: under provision (insufficient number of responses) and honesty (accuracy of subjective data). Weaver & Prelec (2013), categorize deviations from honest reporting of subjective beliefs in the following way: Intentional deception, carelessness and inauthenticity. Firstly, intentional deception relates to a subjects' deliberate act of untruthfully reporting their private signal in a survey question for personal gain, real or psychological, (Suziedelyte & Johar 2013) or due to motives such as low motivation, or confidentiality concerns (Beatty et. al 1998). Secondly, carelessness in question response could stem from lack or insufficiency of task-related incentivization (Smith & Walker 1993). Finally, inauthenticity of subjects' responses results from the presence of biases stemming from social-, environmental influences or cognitive heuristics (Weaver & Prelec 2013). Several biases regarding subjective beliefs are frequently discussed in the literature, i.e. social desirability (Krumpal 2013), anchoring of responses toward the centre of the scale (Simonson & Tversky 1992) and non-response bias due to the sensitive nature of the question (Johnson & DeLamater 1976). Furthermore, framing effects (Tversky & Kahneman 1986), order of questions and response options (Krosnick & Alwin 1987) and consistency seeking behaviour (Johansson-Stenman & Svedsäter 2008; Falk & Zimmerman 2013) have all been found to significantly influence the results of subjective data collection. Hence, to retain research integrity, scholars require incentive compatible methods for the elicitation of subjective truths.

In response to the challenges of subjective truth elicitation, Economists have developed new incentive compatible mechanisms aiming to induce more truthful reporting in subjects' beliefs and attitudes. Prelec (2004) proposes the Bayesian-Truth-Serum (BTS) as an incentive compatible method for belief elicitation in surveys. The BTS mechanism is based on the principle of asking subjects two simple questions (Prelec 2004). First, a binary response to a subjective question such as "Do you believe humanity will have established a self-sustaining Mars colony by 2070?". Second, subjects need to provide their estimate on the fraction of people in accordance with the statement. The information score is used to calculate each subjects' payoffs on the basis of her response to the binary statement against the empirical distribution of all respondents. Subjects whose answers are backed by a larger fraction than collectively predicted are assigned high information scores and vice versa (Weaver & Prelec 2013). Essentially, the 'surprisingly common' principle of the BTS mechanism resides on the Bayesian argument that individuals tend to underestimate the true frequencies of others' opinions (Prelec 2004).

Baillon (2017) introduces another method derived from Bayesian reasoning, namely Bayesian Markets, for the elicitation of private beliefs in the market domain. Bayesian market design due to differences in agents' beliefs, and their influence on expected valuation and subsequent ability to collect the expected positive payoff in a market setting, makes truth-telling a Bayesian-Nash Equilibrium (Baillon 2017). Miller, Resnik & Zeckhauser (2005) present the Peer-Prediction method which uses responses to subjective attitude questions for updating an underlying probability distribution. The incentive scheme scores a rater's response based on a comparison between the actual response and the likelihood assigned to each response using the updated probability distribution (Miller, Resnik & Zeckhauser 2005).

A key challenge with truth-telling incentives is to determine whether these methods indeed enhance data quality. Due to the unverifiable nature of subjective truths researchers have taken different approaches to create benchmarks against which to measure truthfulness. In testing the BTS, John, Loewenstein & Prelec (2012) survey questionable research practices among psychology scholars. Their experiment finds BTS incentives to have a positive effect on self-admission rates for highly sensitive research malpractices, against a non-incentivized control group. A different approach by Weaver & Prelec (2013) finds BTS incentives to outweigh competing incentives to exaggerate recognition of brands in a survey. Baillon, Bleichrodt & Granic (2018) implement a default bias in the user-interface of their survey tool to create a cost to respondents stating their correct belief. The experiment finds BTS incentivization to improve data quality in the presence of the default bias. However, BTS incentives have no effect on participants effort in the absence of the induced bias (Baillon, Bleichrodt & Granic 2018). Furthermore, learning effects have been found to increase participants payoffs in sequential game settings for truth-telling incentives such as the BTS (Weaver & Prelec 2013) and the Peer Prediction Method (Gao, Mao, Chen & Adams 2014). Finally, Bayesian Markets have been found to induce truth-telling when agents believe other players to behave truthfully (Xu 2016).

However, these methods also have their drawbacks, as they might be intransparent in the calculation of payoffs and hard to explain to subjects (Weaver & Prelec 2013), or prone to bias in the presence of pluralistic beliefs (Baillon 2017). Any implementation of the BTS comes with trade-offs to the researcher because of its strong experimenter demand effects and the complexities of explaining the information score to participants (Barrage & Lee 2010). Inherent to the Bayesian reasoning, Prelec (2004) assumes a common prior among subjects, implying that subjects utilize Bayesian reasoning to update their individual beliefs, from a commonly held underlying probability distribution. Similarly, the Peer-Prediction method heavily relies on the common prior assumption in both the agents and mechanism (Witkowski & Parkes 2012). Baillon's (2017) method of Bayesian Markets for elicitation of private information allows for deviations from the common prior assumption which are likely to occur in a real-world data setting. Xu (2016) finds Bayesian markets to become less effective when there is increasing noise in participants beliefs regarding the other agent's truthfulness. Additionally, Bayesian Markets are difficult to implement for sequential elicitation of signals (Baillon 2017), as trends with herd behaviour bubbles might arise when individuals are influenced by the market's historical performance (Xu 2016).

Given the complexities in hand with these methods the literature provides effort to refine these, for instance allowing for deviations from the common prior assumption (Witkowski & Parkes 2012) or adjusting the mechanism for smaller populations and non-binary signals (Radanovic & Faltings 2013). Baillon & Xu (2019) offer a simplified Bayesian truth-telling incentive for the elicitation of private binary signals, based on betting as an elicitor of unverifiable truths. Additionally, the method distinguishes itself from the previously mentioned by omitting the common prior assumption among agents (Baillon & Xu 2019). Instead, TFTB is based on an assumption of common interpretation for a collection of items within agents, thus allowing for heterogeneity in agents priors and perceived signal technologies. In a TFTB setting, an agent is expected to truthfully reveal her private signal when she is faced with the choice of betting:

Top Bet: Item X' Score > Random item Y's (unknown to the agent) Score

Flop Bet: Item X' Score < Random item Y's (unknown to the agent) Score

Similarly, a **Target bet** requires the subject to bet whether item X or random item Y has a score above a specified threshold (Baillon & Xu 2019). TFTB requires a rated collection of items that are comparable, a decision maker who received a private signal for at least one item in the collection, and a reward linked to the outcome of the bettor's choice. While other truth-inducing methods rely on posing two separate questions for signal elicitation, TFTB voids this requirement by betting on the relative performance of an item compared to another random item (Baillon & Xu 2019). The simplicity and familiarity of betting should make TFTB an intuitive truth inducing method for binary signals, with less strong assumptions and fewer overall requirements.

Utilizing the TFTB method to improve course evaluations at the Erasmus School of Economics could prove valuable in producing better quality data and hence improving decision making upon the received feedback. Traditional rating or survey mechanisms as currently operationalized in the ESE course evaluations have much room for bias since responders may fear retaliation to their rating and hence abstain from revealing their private signal truthfully (Miller, Resnik & Zeckhauser 2005). Other studies find SETs to be an opportunity to vent negative emotions following discontent with the teaching or examination (Marlin 1987). Further, the use, reliability and validity of student evaluations of teaching remains a disputed topic in research on higher education (see e.g. Marsh 1984; Shelvin, Banyard, Davies & Griffiths 2000). Against student perception, SETs are a decision tool that is utilized for important personnel decisions at University faculties (Marlin 1987). As a consequence, it is important to investigate the influence of truth-telling incentives on data quality of SETs.

2.1 Research Question & Hypotheses

From the above argumentation, the following research question was derived to assess the truth-telling incentivization via betting on SET outcomes:

"Is incentivization using TFTB effective in reducing default bias in the domain of student course evaluations?"

In order to estimate the effect of the proposed TFTB method, a comparison of data quality between task-related (TFTB) and non-task related incentivization (Random Lottery Incentive) is performed. Specifically, the interest lies in indicators of data quality such as reduction in default bias, response time, drop-out rates and subjective valuations of bets as proxies for effort and strength of signal respectively. Assessing exerted effort in economic experiments is a common measure of interest in discussing the effects of incentivization (Charness, Gneezy & Henderson 2018). Moreover, Theory of principal agent models (Lazear 1986) and the empirical literature (Deci 1971) assume that task-related incentivization yields more effort in agents. An implementation of default bias in reference to Baillon, Bleichrodt and Granic (2018) creates a cost to the respondent for revealing subjective truth if it deviates from the preselected option. In the presence of default bias, we expect TFTB incentivization to increase the effort of participants in comparison to the effort under non-task related incentivization. Therefore, if TFTB incentives indeed increase participants effort, the effects should be observable as differences in frequencies of the SET questions due to the default bias. The incentivization effects on effort of the TFTB Method are evaluated with the help of the hypotheses below:

H1a: Top/Flop betting incentivization reduces default bias as compared to non-task related incentivization.
H1b: Target betting incentivization reduces default bias as compared to non-task related incentivization.

Further analysis on the exerted effort of participants relates to response time, the amount of time an individual took to answer the survey. Response time can be a helpful tool in identifying low effort and carelessness among subjects (Spiliopoulos & Ortmann 2018). Although Rubinstein (2007) advises against the use of response time in small samples it is still implemented as an exploratory proxy for effort due to neglectable implementation cost.[2] Thus, provided that TFTB increases participants effort, we expect the following to hold:

H2a: Response time for Top/Flop betting incentivization is higher than non-task related incentivization.
H2b: Response time for Target betting incentivization is higher than non-task related incentivization.

[2] It is trivial to include response time measurement in the survey tool Qualtrics

Since response rates are also assumed to improve under incentivization (Porter & Whitcomb 2003; Laguilles, Williams & Saunders 2011), we test whether TFTB incentivization is sufficient to result in less students dropping out of the survey. If the TFTB incentivization motivates students to complete the survey more than non-task related incentivization, we would observe significant differences in drop-out rates between the two. The third hypothesis we derive from this argument states as follows:

H3a: Drop-out rates for Top/Flop betting incentivization are lower than non-task related incentivization.
H3b: Drop-out rates for Target betting incentivization are lower than non-task related incentivization.

Finally, we assess the strength of signals for all Top/Flop and Target bets using the BDM-valuation method as proxy (Becker, Degroot & Marshak 1964). Subjects WTA is elicited ranging between 1 and 10 tickets. Since subjects are offered two bets for each choice they make, we expect an implicit probabilistic evaluation of these bets. The subject is expected to choose the bet which she believes is most likely to yield the reward. Therefore, if subjects behave consistently and make a probabilistic comparison of their choice options, the valuation of chosen bets should range between 5 and 10 tickets in the majority of WTAs.

H4a: A significant fraction of Top/Flop bets is valued with at least 5 tickets.
H4b: A significant fraction of Target bets is valued with at least 5 tickets.

3 Experimental Design

3.1 Top/Flop and Target Betting - Discussing the Assumptions

In order to show the suitability of TFTB incentives for student evaluations of teaching, we discuss the assumptions of the method. First, Assumption 1 of the TFTB method states how information on an item score is related to the likelihood of generating a positive signal (Baillon & Xu 2019). In the context of course evaluations, assume an agent learns the score of the course Mathematics 1 to be 7.2 out of 10. She uses this information to update her subjective probability for a positive signal of Mathematics 1 upwards. Alternatively, say instead she learns the score of Mathematics 2 to be 7.2. She can also use this information to update her probability of liking Mathematics 1, but under Assumption 1 the score of Mathematics 1 will have a stronger positive correlation on forming a positive signal than the score of Mathematics 2.

Secondly, TFTB requires the assumption that prior to receiving a private signal the subject treats the collection of items equivalently in her expectation of scores (Baillon & Xu 2019). Assumption 2 is the varied form of the common prior assumption as outlined in the related literature. In our experimental design, this assumption requires the categorization of evaluated courses into Mathematics and Economics courses. This separation was necessary, because students could have different perceptions on the (expected) outcomes of scores for each category, and hence would hold different priors for these.

Thirdly, Assumption 3 states that the scores of the collection are conditionally independent, implying that changes in posterior probability due to information about one score are invariant to information about other scores. Therefore, using the argumentation from Assumption 1, an agent's posterior probability of a positive signal after learning the score of Mathematics 1 does not change when additionally, information on the score of Mathematics 2 is revealed.

The fourth assumption of TFTB states that subjects behave according to probabilistic sophistication in reference to Machina & Schmeidlers (1992) definition. Under this assumption, a subject can behave risk-neutral or seeking under expected utility, however the method allows for deviations from expected utility theory as long as the probability weighting transformation is strictly increasing (Baillon & Xu 2019). In essence, Assumption 4 states that an agent will choose the bet that maximizes her subjective probability of obtaining the reward. Our testing of *H4* will give an indication whether the assumption holds in the obtained student sample.

Lastly, Assumption 5 of the TFTB method requires common knowledge in agents that Assumptions 1-4 hold. This Assumption allows subjects to have different individual prior and posterior probabilities of generating a positive signal. Further, the requirement of assumption 5 is that agents agree on the relationships of prior and posterior probabilities regarding the scores of the collection of items. In practice, his implies that subjects agree on the informational properties of SET scores regarding the probability of liking the course. For example, all subjects agree that their posterior probability of liking Mathematics 1 is higher after learning Mathematics 1 SET score as compared to learning the SET score of Mathematics 2.

3.2 Subjects and Experimental Procedure

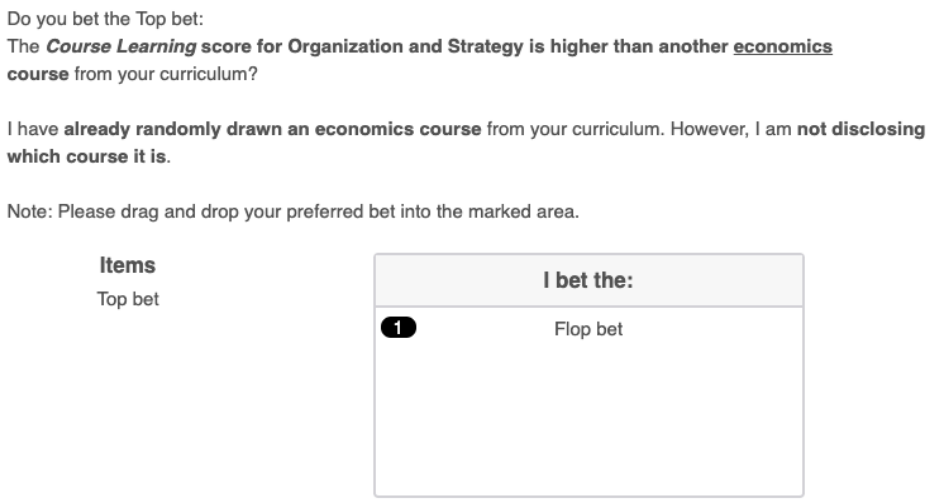
For the experiment, two groups of students from Erasmus School of Economics' International Bachelor of Economics & Business Economics (and its Dutch counterpart: Economie en Bedrijfseconomie) were invited to take part in an online course evaluation via announcements in Facebook groups and the Universities' online learning platform CANVAS. The experiment was conducted in two stages. The first being a condensed version of current student course evaluations, and the second being an experimental variation with an implementation of the Top/Flop and Target betting method.

The first stage of the experiment consisted of an incentivized course evaluation on courses in years 1 and 2 of the current curriculum. For this stage, year 3 students were invited to participate in order to guarantee they had experienced the selection of 9 courses during their studies. Usually, course evaluations pose between 20 and 30 questions, depending on the number of Professors and Tutors involved in the course.

The survey contained the typical questions from SETs regarding learning achievement (5-Point Likert scale), course organization (5-Point Likert scale) and overall grading of the course (1-10 grading scale). As compensation, the students were provided with the opportunity to enrol in a Binary Lottery where the number of tickets depended on the number of courses they chose to evaluate. Subjects could opt for evaluating five or nine courses. When subjects opted to rate 9 courses (n=30) they were presented with 30 questions (26-course evaluation and 4 demographic questions), with each course evaluation block presented in random order, yielding eleven lottery tickets upon completion. Similarly, Subjects that evaluated five courses (n=6), were presented with evaluation questions for five courses in random order, receiving five lottery tickets upon completion. Preliminary to answering questions on each course, participants were presented with a memory-refresher page regarding the Block in which they had taken a course, and a short overview of the main materials covered in the lectures. In total, the first wave of the experiment hat 37 clicks, 28 finished responses and 20 respondents that opted into the Binary Lottery incentive for a chance to win a 20 Euro Amazon Coupon. Of the 28 finished responses 61% came from male participants, the obtained ratio of male to female students is not significantly different from the officially stated ratio of 50:50. The collection of course evaluation data from year 3 students served as means to construct the bets for stage two based on data which resembles outcomes of previous ESE course evaluations. Thus, stage one ensured the comparable collection of items necessary under Assumption 2 of the TFTB method.

3.3 Stage 2 - Experimental Treatments

Stage two of the experiment tests the TFTB method against several indicators of data quality. Each type of bet (Top/Flop or Target) refers to a single treatment and control group. After identification of their study year, subjects of year one (n=36) were stratified into the Top/Flop treatment-control sample, and year two subjects (n=79) into the Target treatment-control sample. To differentiate treatment and control groups in terms of data quality, each variation of the survey included an experimenter induced default bias similar to Baillon, Bleichrodt & Granic's (2018) implementation. The default bias (Figure 1), with the same option being preselected for all participants, created a cost to correctly stating one's beliefs. If a subject was of a different opinion than the default, she had to remove the default option and select her preference via drag and drop. This default bias also provided an opportunity for a lazy participant to shirk by advancing quicker or made it easier for those who were in agreement with the default option.



Do you bet the Top bet:
The ***Course Learning* score for Organization and Strategy is higher than another economics course** from your curriculum?

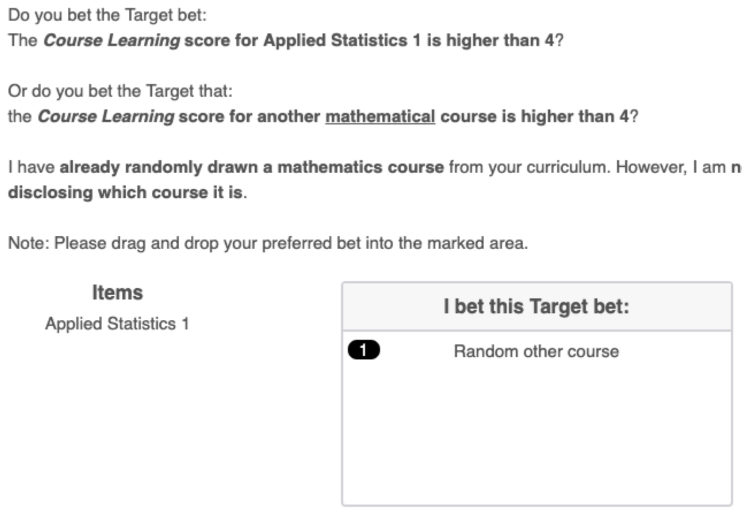
I have **already randomly drawn an economics course** from your curriculum. However, I am **not disclosing which course it is**.

Note: Please drag and drop your preferred bet into the marked area.

Figure 1: Example of operationalized default bias for a Top/Flop betting choice

Both control and treatment groups were first shown the same instructions informing them they were about to take part in an incentivized course evaluation. After stratifying into treatments, participants received instructions (see Appendix 1 & 2 for TFTB instructions) on the procedure and payoff structure based on treatment. Before each course specific question block, participants received the memory-refresher pages of stage 1. In the control groups, subjects had flat-fee incentivization by a Random Lottery Incentive (10 tickets for a 20 Euro Amazon coupon lottery) and were asked to evaluate two Economics and one Mathematics course of their curriculum. The questions in the online survey asked to evaluate course organization, learning outcomes, relevance and understandability of the materials and overall course liking. In order to make these questions similar to the betting methods in the experimental treatments, the scale for course organization, learning outcomes, and relevance and understandability was varied from a 5-point Likert to a binary Yes/No scale since TFTB requires the elicitation of binary signals. Moreover, to have the same number of questions per course in the treatment and control groups, a fourth question on overall course liking was elicited on a 10-point scale similar to the valuation task in both treatment groups.

In the treatment groups, subjects received task-related incentivization through a Binary Lottery incentive utilizing the TFTB method. In both treatment groups, subjects evaluated two Economics and one Mathematics course from their curriculum. Each course was evaluated by betting on course learning, organization and overall grade followed by a BDM-valuation task (see Appendix 3). Subjects received ten tickets for completion of the survey and ten additional tickets per bet won. Treatment group 1, received these questions with the framing in Figure 1, while for Treatment group 2 the framing was according to Figure 2.



Do you bet the Target bet:
The ***Course Learning* score for Applied Statistics 1 is higher than 4**?

Or do you bet the Target that:
the ***Course Learning* score for another mathematical course is higher than 4**?

I have **already randomly drawn a mathematics course** from your curriculum. However, I am **not disclosing which course it is**.

Note: Please drag and drop your preferred bet into the marked area.

Figure 2: Example of operationalized default bias for a Target betting choice

After placing three bets on a given course, subjects were asked to provide a valuation of their bet in terms of lottery tickets deploying the BDM valuation method. For this part of the survey, subjects were instructed to provide their Willingness-to-Accept in a selling scenario with a price already drawn by the experimenter, but not revealed to subjects.

4 Results

Before we proceed with hypotheses testing, we provide a table with the variables used in our analysis (Table 1). Furthermore, we present the outcomes of the student evaluation for stage 1 (Table 2). The outcomes of the first stage of our SET are used to construct the bets of the second stage. Generally, we observe that Economics and Mathematics courses scored similar on average in course learnings (Econ: 3.984; Math: 3.991). Whereas Mathematics courses scored higher on average in the categories of course organization (Econ: 3.803; Math: 4.067) and Overall grade (Econ: 7.231; Math: 7.782). These outcomes affected our determination of bets in the Target treatment. Thus, we set Targets around the mean for Economics courses at 4, 3.8 and 7 and 4, 4 and 7.7 for Math courses in the order of categories from Table 2. This is to ensure that a Target bet cannot win with certainty.

Variable	Description
Demographics:	
Gender	Binary gender identification, 1= Male, 0=Female
Cohort	Identification of student cohort, 0= International Bachelor Economics & Business Economics, 1= Economie en Bedrijfseconomie, 2=other
Year	Identification of study year, 1= first year of studies, 2=second year of studies
Stage 1:	
Incentive Choice	Indicator of number of evaluated courses based on choice of incentive, 1=5 courses (5 tickets), 2=9 courses (11 tickets)
Course Learnings	Opinion on "I have learned a lot during this course", 5-point Likert scale, 1=Strongly Disagree, 5=Strongly Agree
Course Organisation	Opinion on "The course was well organized", 5-point Likert scale, 1=Strongly Disagree, 5=Strongly Agree
Course Grade	Subjective grade on overall course liking, 1-10 scale
Stage 2:	
Treatment	Indicator of Treatment, 1=TFTB incentivization, 0=Control group (Binary Lottery Incentive)
Signal	Elicited signal from subject, 1=positive, 0=negative (default option)
Course #	Indicator of course, (1-5) courses evaluated in stage 2, 1=Economics#1, 2=Economics#2, 3=Economics#3, 4=Economics#4, 5=Mathematics#1, 6=Economics#5, 7=Economics#6, 8=Mathematics#2, 9=Mathematics#3
Item response time	Time in seconds until submission of evaluation question
Order	Indication of question order as presented to subject, values 1-9
Evaluation Category	Referring to the evaluation criteria, 1=Course Learnings, 2=Course Organisation, 3=Overall Grade/Relevance
BDM-Valuation	Valuation of choice betting choice for evaluation category 3, 1-10 tickets
Finished	Indicator for finished response, 1=complete 0=incomplete

Table 1: Description of variables used in the analysis of stage 1 and 2

	(1)	(2)	(3)	(4)	(5)
Course & Category	**N**	**mean**	**sd**	**min**	**max**
Economics #1					
Learnings	28	**3.464**	1.105	1	5
Organisation	28	**3.750**	1.110	1	5
Overall Grade	28	**6.786**	1.771	3	9
Economics #2					
Learnings	27	**4.333**	0.961	2	5
Organisation	27	**3.926**	0.874	2	5
Overall Grade	27	**7.852**	1.350	5	10
Economics #3					
Learnings	28	**3.643**	1.026	1	5
Organisation	28	**3.643**	1.129	1	5
Overall Grade	28	**6.821**	1.611	3	10
Economics #4					
Learnings	28	**4.321**	0.905	2	5
Organisation	28	**4.286**	0.854	2	5
Overall Grade	28	**7.821**	1.611	2	10
Economics #5					
Learnings	28	**4.107**	1.066	2	5
Organisation	28	**3.893**	1.100	1	5
Overall Grade	28	**7.536**	1.503	4	10
Economics #6					
Learnings	28	**4.036**	0.962	2	5
Organisation	28	**3.321**	1.124	1	5
Overall Grade	28	**6.571**	1.834	2	9
Mathematics #1					
Learnings	26	**4.423**	0.758	3	5
Organisation	26	**3.962**	0.824	2	5
Overall Grade	26	**7.692**	1.463	4	10
Mathematics #2					
Learnings	29	**3.552**	1.055	1	5
Organisation	29	**4.103**	0.772	2	5
Overall Grade	29	**7.862**	1.026	6	10
Mathematics #3					
Learnings	29	**4**	0.926	2	5
Organisation	29	**4.138**	0.915	1	5
Overall Grade	29	**7.793**	1.424	3	10

Table 2: Stage 1 course evaluation outcomes for all 9 courses, descriptive statistics

4.1 Top/Flop Betting Results

4.1.1 Payment Mechanism & Evaluation Outcomes

Next, we present the outcomes of the evaluation for the Top/Flop treatment and the fraction of bets won in each course category, the averages in Table 3 do not contain the signals of the control group. Table 3 shows the fraction of subjects that rejected the default for both Economics courses and the Mathematics course which were evaluated in the Top/Flop Treatment. We observe that subjects' responses varied between courses, and across the different evaluation categories such that the obtained frequencies of positive signals range between 9% and 80%. In the Top/Flop treatment group, the fraction positive signal for economics courses was 45.5% for course learnings, 22.7% for course organization and 58.2% for overall grade. We can already observe that the default bias was effective in reducing the number of positive signals. Alternatively, Top/Flop betting could have induced more truthful responding resulting in students revealing their true (more negative) attitudes toward certain courses.

Course & Category	(1) N	(2) mean	(3) sd
Economics #1			
Learnings bet	**11**	**0.273**	**0.467**
Win percentage of bets	11	0.727	0.467
Organisation bet	**11**	**0.364**	**0.505**
Win percentage of bets	11	0.818	0.405
Overall grade bet	**10**	**0.800**	**0.422**
Win percentage of bets	10	0.100	0.316
Economics #2			
Learnings bet	**11**	**0.636**	**0.505**
Win percentage of bets	11	0.636	0.505
Organisation bet	**11**	**0.0909**	**0.302**
Win percentage of bets	11	0.182	0.405
Overall grade bet	**11**	**0.364**	**0.505**
Win percentage of bets	11	0.364	0.505
Mathematics #1			
Learnings bet	**9**	**0.667**	**0.500**
Win percentage of bets	9	0.667	0.500
Organisation bet	**9**	**0.444**	**0.527**
Win percentage of bets	9	0.556	0.527
Overall grade bet	**9**	**0.444**	**0.527**
Win percentage of bets	9	0.556	0.527
Total Tickets	18	32.33	27.29

Table 3: Stage 2, fraction of positive signals & Win rates of bets in the Top/Flop Treatment course evaluation

In order to determine the winner of the Top/Flop betting lottery, we calculated the number of tickets per subject as follows. For each individual bet, we draw the random course and its category score linked to the Flop bet. Then based on the choice of Top or Flop, we determine whether the subject won the bet, receiving 10 tickets for each correct bet. The mechanism seemed to work quite well, there is variation with regards to the fraction of successful bets between 10% and 82%. Furthermore, for each bet in the category of 'Overall Grade', we resolve the selling scenario of the BDM valuation task. Based on the randomly drawn price, we sell each bet valued below the price, endowing subjects with the number of tickets stated by the price. In case subjects did not sell their bet, they receive the outcome of their bet (10 or 0 tickets). From the 8 complete responses in the Top/Flop treatment, 6 opted into the lottery. For these, we issued 378 tickets with an average of 63 tickets per subject, drawing a random number between 1 and 378 to determine the winning ticket. The winner was then notified via the email entered at the end of the survey.

4.1.2 Default Bias

To analyse the effect of the Top/Flop betting incentivization on data quality, we first test whether the obtained distributions of signals between treatment and control are equal. From the observed evaluation outcomes in Table 3, we already note that the frequencies of three evaluation questions in the treatment group tend towards the default (negative signal regarding the course). First, we run a Fisher exact test on the aggregated signals from all courses and categories, excluding the category of overall grade (treatment) and relevance of materials (control) to void these confounding factors. Then, we run two Fisher's exact tests per course, for course learning and course organization respectively (Results in Table 4).

Course & Category	Treatment	FREQ Signal (+)	FREQ Signal (-)	Total	one-sided p-value
Pooled Signals					
	Control	48	39	**87**	**0.052**
	Top/Flop	25	37	**62**	
	Total	**73**	**76**		
Economics#1					
Learnings	Control	6	9	**15**	**0.402**
	Top/Flop	3	8	**11**	
	Total	**9**	**17**		
Organisation	Control	9	6	**15**	**0.214**
	Top/Flop	4	7	**11**	
	Total	**13**	**13**		
Economics#2					
Learnings	Control	9	5	**14**	**0.648**
	Top/Flop	7	4	**11**	
	Total	**16**	**9**		
Organisation	Control	7	7	**14**	**0.038**
	Top/Flop	1	10	**11**	
	Total	**8**	**17**		
Mathematics #1					
Learnings	Control	8	7	**15**	**0.418**
	Top/Flop	6	3	**9**	
	Total	**14**	**10**		
Organisation	Control	9	5	**13**	**0.306**
	Top/Flop	4	5	**9**	
	Total	**13**	**10**		

Table 4: Fischer exact test outcomes for 7 distributions from course evaluation stage 2, Top/Flop bets

The obtained results from the seven Fisher exact tests yield insignificant results at the 5% significance level except for one, indicating that signal frequencies in treatment and control come from similar distributions. However, we find significantly different distributions between control and treatment group for Econ Course #2 (p=0.038) in the course organization category. Frequencies in this category are dominated by negative signals for the treatment group (10 Flop bets vs 1 Top Bet). This result either indicates that the treatment group has truly different signals as compared to the control group (7 positive vs 7 negative signals), or that the default bias was effective in providing an opportunity to shirk in the treatment.

Hence, based on the results of the Fisher exact test on the aggregate distribution of signals, we find no evidence for a reduction of default bias under Top/Flop betting incentivization. Therefore, we do not accept Hypothesis 1a. This result implies that Top/Flop betting does not outperform a Random Lottery Incentive in the presence of a one-directional default bias. However, we need to note that the higher frequency of negative signals in the treatment group relative to its control results in an almost significantly different distribution toward the default. This outcome does not match our hypothesized impact of the default, indicating that either the method on average induced more truthful responding towards negative signals, or that the incentivization did not sufficiently motivate our subjects in the treatment group.

4.1.3 Drop-out Rates

Another indicator of data quality we analyse is the number of subjects dropping out of the survey. For this reason, we tested whether the proportions of incomplete responses are larger in the control relative to the treatment group. Large sample conditions (np and n(1-p) > 5) for the binary T-test are not met for the observed distribution, hence we only perform a Fisher Exact test for equality of distributions.

Treatment	Complete	Incomplete	Total	one-sided p-value
Control	14	4	**18**	**0.043**
Top/Flop	8	10	18	
Total	**22**	**14**		

Table 5: Fischer exact test outcome for drop-out distributions

The result of the Fisher exact test in Table 5 (p=0.043) indicates that, against expectations, the treatment suffered from significantly higher attrition relative to the control group. In considering this difference in drop-outs we refer to two possible causes, insufficient incentives and unfamiliarity with the methods. These could have impacted a subject's decision to drop-out. In turn lingering the motivation of subjects to finish the survey as they proceeded. This complexity could have raised the subjective cost beyond the perceived benefit of incentivization for first-year students. Among the ten drop-out subjects in the treatment group, we find that two dropped out during treatment instructions and the remaining eight without a distinct pattern over the course of the survey. On the contrary the control group subjects faced the familiar questions from previous evaluations requiring much less effort. In conclusion to this subsection, we find contrary evidence for Top/Flop incentivization for H3a and reject the hypothesis.

4.1.4 Response Time

Treatment	N	Mean (Total Response Time)	Std. Err.	Std. Dev.	CI (95%)	
Control	18	**94.167**	**11.883**	**50.419**	**69.094**	**119.239**
Top/Flop	**18**	**195.667**	**33.853**	**143.627**	**124.243**	**267.091**

Table 6: Average total response time by Top/Flop Treatment and Control

In analysing the effect of Top/Flop target betting on response time, we look at the average total response time for both treatment and control (Table 6). We find that subjects in the Top/Flop Treatment spend on average significantly more time (p= 0.04) with 195 seconds in the survey relative to the control group with a mean of 94 seconds. Considering this result includes incomplete responses, we are encouraged that subjects indeed spend more effort on the evaluation tasks in the Top/Flop treatment.

Next, we analyse participants effort in the course evaluation measured by item response time. By item response time we refer to the amount of time it took a participant to submit an answer to an evaluation question. For the analysis we use a logarithmic transformation of the variable item response time to approximate a normal distribution. Further, we run a pooled-OLS regression on response time with clustered standard errors on participant ID. Our variables of interest are the Top/Flop treatment, signal type, an interaction effect of these two and an indicator of the first evaluation item (bet or attitude question that was first shown to the participant). Further, we control for order fixed effects, finished responses, student cohort, and gender. Additionally, we include an interaction term for the order of questions and treatment, testing whether learning effects regarding the Top/Flop method are observable. We fit three models, one for all three courses under evaluation, and two models restricting the sample to Economics and Mathematics courses respectively. All three models' results are displayed in Table 7.

Dependent Var: ln(Response Time)	(1) All courses	(2) Econ courses	(3) Math course
Top/Flop Treatment	1.645***	1.551***	1.689**
	(0.396)	(0.409)	(0.604)
Positive Signal	0.995***	1.086***	0.801*
	(0.332)	(0.290)	(0.433)
Signal-Treatment Interaction	-0.488	-0.526	-0.540
	(0.367)	(0.352)	(0.485)
First evaluation question	0.961***	1.090***	0.917***
	(0.164)	(0.230)	(0.243)
Finished response	0.375	0.236	1.275
	(0.298)	(0.290)	(0.760)
Order-Treatment Interaction	-0.0764**	-0.0483	-0.0967
	(0.0348)	(0.0558)	(0.0727)
Constant	0.450	0.547	-0.396
	(0.469)	(0.407)	(0.895)
Observations	222	151	71
R-squared	0.516	0.552	0.505
Clustered on	ID	ID	ID
Demographic Controls	YES	YES	YES
Order Controls	YES	YES	YES

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 7: pooled-OLS regression estimates of Top/Flop Treatment-Control sample

Our implementation of a default bias was effective in increasing the cost of subjects switching to the positive signal. The coefficient of Signal in model (1) highlights this cost as subjects in both groups spent on average 100% more seconds when deviating from the default. This effect persists in model (2), while decreasing in size and significance for model (3). Of course, the default also provided a convenience for subjects whose signal was the preselected default, or those who were unmotivated and rushed through the survey thus shortening response time. Furthermore, the results indicate that subjects in the Top/Flop treatment according to model (1), on average spent 165% more seconds per question when submitting a negative signal, and 215% more seconds when submitting a positive signal. This effect is robust in the estimation of both subsamples for model (2) and (3). Therefore, the Top/Flop incentivization was effective in motivating subjects to respond with careful consideration and spend more time per question relative to the control group. As we hypothesized, the Top/Flop treatment induced significantly higher effort in subjects measured by response time, leading us to accept H2a. Generally, keeping in mind the substantial attrition in both groups, subjects seemed to enter the survey with high initial motivation spending on average 96% more seconds on the first question item presented to them. In the treatment group, question order significantly affected response time in model (1), as subjects on average spent 7,6% less time per additional question item presented to them. This either implies a learning effect with regard to the Top/Flop method, or it represents a decrease in motivation as the survey progressed.

4.1.5 Bet Valuation

In order to determine whether the probabilistic sophistication assumption of the TFTB method holds, we analyse the WTA for the Overall Grade bets of subjects in the Top/Flop treatment. If subjects are probabilistically sophisticated, they will choose the bet with the higher subjective probability of obtaining the reward. Therefore, provided that subjects don't guess their valuation in the BDM task, the elicitation should reflect a subjective probability of at least 50%. Under the acclaimed risk neutral BDM elicitation, the stated ticket price should therefore reflect a subject's subjective probability of winning the bet. For the analysis we pool the outcomes of the three BDM valuation tasks in the Top/Flop treatment. Subsequently, we determine the fraction of bets valued at or above 5 tickets. We run a binomial test with the fraction of bets valued with at least 5 tickets against a hypothetical distribution (p=0.5) which would pertain if subjects randomly guessed.

	N	Observed (>=5)	Expected (>=5)	Assumed p	Observed P	P (k>=28)	P (k<=4 or k>=28)
Top/Flop Valuations	32	28	**16**	**0.5**	**0.875**	**0.00**	**0.00**

Table 8: results of binomial test of probability for BDM-valuations in the Top/Flop treatment

Our result indicates that the observed distribution (87.5% of bets valued above 5) is significantly different (two-tailed p=0.000019) from the hypothesized distribution of 50%. This finding implies that subjects in the Top/Flop treatment behave according to probabilistic sophistication and choose the bets according to their subjective probability. Therefore, we find evidence for H4a showing that subjects in a Top/Flop bet setting make a conscious choice in line with the probabilistic sophistication assumption in the majority of cases.

4.2 Target Betting results

4.2.1 Payment Mechanism

For the analysis of Target bets efficacy in improving data quality, we proceed in the same way as with the previous treatment. Table 9 shows the win rates and the fraction of signals per course and evaluation category for the Target treatment. In the Target treatment, the mechanism yielded winning rates between 33% and 76%. Further, we observe that the average fraction of positive signals for economics courses is 41,7% in the category of course learning, 40,4% in course organisation and 38.3% in overall grade. We note that the default seems to be rejected more frequently than in the Top/Flop Treatment. This could be a possible indication that the method indeed induces subjects to consider their answer before submitting a response. It is notable that the fraction of positive signals for the overlapping course Mathematics #1 is different in the Target treatment relative to the Top/Flop treatment (Table 9). This difference could be due to the increased number of observations or a different perception of year two students at the end of their second year of studies.

Course & Category	**(1) N**	**(2) mean**	**(3) sd**
Economics #3			
Learning bets	**18**	**0.278**	**0.444**
Win percentage of bets	18	0.450	0.510
Organisation bet	**18**	**0.278**	**0.461**
Win percentage of bets	18	0.370	0.492
Overall grade bet	**17**	**0.412**	**0.507**
Win percentage of bets	17	0.529	0.514
Economics #4			
Learning bets	**18**	**0.556**	**0.511**
Win percentage of bets	18	0.556	0.511
Organisation bet	**17**	**0.529**	**0.514**
Win percentage of bets	17	0.706	0.470
Overall grade bet	**17**	**0.353**	**0.493**
Win percentage of bets	17	0.588	0.507
Mathematics #1			
Learning bets	**17**	**0.765**	**0.437**
Win percentage of bets	17	0.765	0.437
Organisation bet	**17**	**0.706**	**0.470**
Win percentage of bets	17	0.294	0.470
Overall grade bet	**17**	**0.765**	**0.437**
Win percentage of bets	17	0.235	0.437
Total Tickets	34	35.06	27.09

Table 9: Stage 2, fraction of positive signals & Win rates of bets in the Target Treatment course evaluation

Furthermore, we calculated the number of tickets per subject based on their betting choice and our specified targets. In the survey, we specified the targets for Economics courses at 4, 3.8 and 7 for the categories of Learning, Organisation and Overall Grade respectively. For the Mathematics course, we set targets at 4, 4 and 7.7 in the same order of categories. The number of tickets per subject was calculated after determining the random other course for all bets and obtaining the respective scores. Moreover, we resolve the selling scenario of the BDM task in the same way as stated above in the Top/Flop treatment. From the 15 complete responses in the Target treatment 13 opted into the lottery. We issued 522 tickets with an average of 40 tickets per subject. The winner was then notified via the email entered at the end of the survey.

4.2.2 Default Bias

In analysing the effect of Target betting on data quality we test whether the observed fractions of signals show significant differences between the treatment and control groups. Due to the higher number of observations in both groups relative to those in Top/Flop betting, we also run a t-test of proportions in addition to the fisher exact test. Both tests are run for the evaluation categories of Learning and Organisation for all three surveyed courses as well as the aggregated sample of all signals.

Course & Category	Treatment	FREQ Signal (+)	FREQ Signal (-)	Total	one-sided p-value (Fischer exact)	Mean diff.	one-sided p-value (t-test)
Pooled Signals							
	Control	116	42	**158**	**0.000**		
	Target	54	53	**107**			**Pr(Z>z)**
	Total	**170**	**95**			**0.230**	**0.000**
Economics#3							
Learnings	Control	23	4	**27**	**0.000**		
	Target	5	13	**18**			**P(Z>z)**
	Total	**28**	**17**			**0.574**	**0.000**
Organisation	Control	23	4	**27**	**0.001**		
	Target	5	13	**18**			**P(Z>z)**
	Total	**28**	**17**			**0.574**	**0.000**
Economics#4							
Learnings	Control	20	6	**26**	**0.122**		
	Target	10	8	**18**			**P(Z>z)**
	Total	**30**	**14**			**0.213**	**0.0673**
Organisation	Control	20	6	**26**	**0.096**		
	Target	9	8	**17**			**P(Z>z)**
	Total	**29**	**14**			**0.239**	**0.0504**
Mathematics #1							
Learnings	Control	16	10	**26**	**0.247**		
	Target	13	4	**17**			**P(Z<z)**
	Total	**29**	**14**			**-0.149**	**0.1535**
Organisation	Control	17	9	**26**	**0.494**		
	Target	12	5	**17**			**P(Z<z)**
	Total	**29**	**14**			**-0.052**	**0.3609**

Table 10: Fischer exact test outcomes for 7 distributions from course evaluation stage 2, Target bets

For Economics Course #3, we observe significant differences in the Learnings and Organisation evaluation sections. Similar to the Top/Flop betting case, these differences are the result of more negative signals in the treatment group. Hence, the hypothesized reduction of default bias is rejected for Economics Course #3. The observed differences could again be a result of truthful responding, or subjects being tempted to go with the default for some of the question items presented. For the remaining two courses we find no significant differences in the proportions and distributions of signals. Generally, the distribution of signals for the control group stays relatively constant across categories and courses. On the contrary we find that within the treatment the distributions of signals shift towards the default option for the Economics Course #3, this could be an indication that subjects in the Treatment thought carefully about their answers before submitting their bet. If we had employed a yeasayer bias in addition to the existing one, we would be able to observe whether these differences persist for each possible bias. Hence based on the results of both tests for the aggregated sample, we conclude that we find no supporting evidence for H1b.

4.2.3 Drop-out Rates

In determining the effect of Target incentivization on drop-out rates, we analyse the proportions of finished responses in both groups. We run a t-test of proportions in addition to the Fisher exact test, since large sample conditions are met.

Treatment	Complete	Incomplete	Total	one-sided p-value (Fischer exact)	Mean diff.	one-sided p-value (t-test)
Control	20	15	**35**	**0.200**		
Target	15	19	**34**			**Pr(Z>z)**
Total	**35**	**34**			**0.130**	**0.139**

Table 11: Fischer exact test outcome for drop-out distributions

Our results suggest that there are no significant differences (p=0.139) in drop-out rates between treatment (55% drop-outs) and control (43% drop-outs). This result is robust in the Fisher exact test (one sided p=0.200) which is more suitable for small samples. Our results imply that both types of incentivization similarly affect students' motivation to finish the course evaluation. Therefore, we find evidence to reject H3b. We conclude that both a Random Lottery incentive and Target betting incentivization produce similar drop-out rates in the context of SET's.

4.2.4 Response Time

Treatment	N	Mean (Total Response Time)	Std. Err.	Std. Dev.	CI (95%)	
Control	**33**	**106.849**	**8.665**	**49.775**	**89.199**	**124.498**
Target	**33**	**174.394**	**19.483**	**111.922**	**134.708**	**214.080**

Table 12: Average total response time by Treatment

Table 12 shows the average response time in the Target treatment and control groups. We find that subjects in the Target treatment on average spend significantly (p=0.001) more time in the survey as compared to the control group. Further, we analyse subject's effort in the subsample of second year students, corresponding to the Target and control treatment. In the same fashion as in the Top Flop section, we analyse subject's effort measured by item response time. To normalize the distribution of the response time variable, we use a logarithmic transformation before running our analysis. We fit a pooled-OLS model with logarithmic transformed item response time as dependent variable and clustered standard errors on participant ID. Our variables of interest are the Target Treatment, elicited signal, an interaction effect between the former, and an indicator of the first evaluation item. Additionally, we include controls of question order, finished responses, student cohort and gender. Lastly, we are interested in observing the effect of order on response time in the treatment, hence including an interaction of order and treatment. Again, we fit three models, one for all three courses under evaluation, and two models restricting the sample to Economics and Mathematics courses respectively. All three models' results are displayed in Table 13.

	(1)	**(2)**	**(3)**
Dependent Var: ln(response time)	**All courses**	**Econ courses**	**Math course**
Target Treatment	1.339***	1.389***	1.378***
	(0.233)	(0.265)	(0.342)
Positive Signal	0.890***	0.924***	0.790**
	(0.224)	(0.270)	(0.304)
Signal-Treatment Interaction	-0.537**	-0.507*	-0.700*
	(0.248)	(0.297)	(0.401)
First Evaluation question	0.782***	0.730***	0.855***
	(0.106)	(0.119)	(0.252)
Finished Response	-0.205	-0.105	-0.452**
	(0.147)	(0.189)	(0.219)
Order-Treatment Interaction	-0.0516*	-0.0648	-0.0191
	(0.0300)	(0.0428)	(0.0624)
Constant	1.309***	1.303***	1.393***
	(0.253)	(0.260)	(0.460)
Observations	394	265	129
R-squared	0.478	0.504	0.456
Clustered on	ID	ID	ID
Demographic Controls	YES	YES	YES
Order Controls	YES	YES	YES

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 13: pooled-OLS regression estimates of Target Treatment-Control sample

The implementation of a default bias significantly increased response time for submitting a positive signal (89% and 169% more seconds spent, conditional on Treatment), creating a cost to subjects whose submitted signal deviates from the preselected. The effect is observable in model (1) and (2) with some insignificant variation in terms of coefficient size, while less significant (at the 5% level) and smaller in size for model (3). Differences in coefficient size and significance could be due to the cognitive engagement in thinking about the random other course, because the curriculum has a stronger focus on economics compared to mathematics. Hence thinking about which courses compose the alternative Target bet takes more effort for economics than mathematics courses. Moreover, model 1 indicates that subjects in the Target treatment relative to the control group, spent on average 134% more seconds when submitting a negative signal, and 169% more seconds when submitting a positive signal. This effect is robust in the estimation of both subsamples for model (2) and (3), while the interaction term of treatment and signal is reduced to the 10% significance level in both subsample regressions. As we hypothesized, the Top/Flop treatment induced significantly higher effort in subjects measured by response time, hence we accept H2b. In the treatment group, question order did not significantly affect response time in three models. This either implies that Target betting is more intuitive than Top/Flop betting, or that subject's motivation stayed constant across the different evaluation questions.

4.2.5 Bet Valuation

In testing the probabilistic sophistication assumption of the TFTB method, we analyse the valuations of the Overall Grade bets in the Target bet treatment. As for Top/Flop bets, subjects are deemed to behave according to probabilistic sophistication if they chose the bet that maximizes their subjective chance of obtaining the reward. Hence, we run a binomial test to determine whether the observed fraction of valuations above 5 tickets deviates from the hypothesized distribution in which subjects guess the valuation. Subjects valuations for all three overall grade bets were pooled before determining the fraction of bets valued with at least 5 tickets.

	N	Observed (>=5)	Expected (>=5)	Assumed p	Observed P	P (k>=48)	P (k<=12 or k>=48)
Target Valuations	60	48	**30**	**0.5**	**0.8**	**0.00**	**0.00**

Table 14: results of binomial test of probability for BDM-valuations in the Top/Flop treatment

The results of the binomial test (two tailed p=0.000003) suggest a significant difference of the observed outcomes of the BDM task (p=0.8) against the hypothesized distribution (p=0.5). This finding implies that also in the Target treatment subjects value their choice of bet consistent with the probabilistic sophistication assumption of the TFTB method, leading us to accept H4b.

5 Balance & Robustness Checks

5.1 Balance Check

In order to determine the validity of our analysis, we run a balance check with regards to both treatment and control groups to ensure no significant differences in characteristics and subsample composition. We test whether the averages of characteristics such as cohort, gender and the composition of finished responses are significantly different. As we found in our Analysis in subsection (Drop-out), Table 13 shows significant differences in the composition of finished responses between the Top/Flop treatment and control groups. The remaining characteristics are balanced, since we find no significant differences between the means of cohort and gender. For the Target treatment-control sample, we find no significant differences for all three variables. Hence, we proceed in our robustness analysis by a) excluding unfinished responses, and b) running supplementary analysis for our regressions.

	Control	**Target Treatment**	**Diff.**
Cohort	**1.857**	**1.853**	**-0.004**
	(0.355)	(0.359)	(0.086)
Male	**0.743**	**0.647**	**-0.096**
	(0.443)	(0.485)	(0.112)
Finished	**0.571**	**0.441**	**-0.130**
	(0.502)	(0.504)	(0.121)
Observations	**35**	**34**	**69**

	Control	**Top/Flop Treatment**	**Diff.**
Cohort	**1.111**	**1.111**	**0**
	-0.323	-0.323	-0.108
Male	**0.5**	**0.5**	**0**
	-0.514	-0.514	-0.171
Finished	**0.778**	**0.444**	**-0.333****
	-0.428	-0.511	-0.157
Observations	**18**	**18**	**36**

Table 15: Balance Checks for Top/Flop and Target treatments & controls

5.2 Robustness Check

5.2.1 Top/Flop Betting

Our results on the reduction in default bias stay robust in the exclusion of unfinished responses except for two results (Appendix 4). The almost significant p-value of the Fischer exact test for the aggregate sample is reduced to p=0.250. For the distribution of Course Learning for Econ Course #3, we find a reduction in p-value (one sided p-value=0.095) making the difference in distributions insignificant at the 5% level. This result implies that upon exclusion of unfinished responses, we now observe no reduction in default bias in the control relative to the treatment group. Therefore, we find further evidence, that a Binary Lottery incentive and Top/Flop target betting produce similar distributions in the presence of a one-directional default bias. A notable observation we make is that the exclusion of unfinished responses leads us to drop 10 negative signals and 3 positive signals across evaluation items in the treatment group, whereas we drop 3 positive signals only for the control group. This structure of unfinished responses is analysed with a binomial test of probability, to investigate whether the dropped-out responses in the treatment have a significant default bias. The result (p(k=>10) = 0.046) of the test (Appendix 5) suggests that the observed distribution of dropped signals is significantly biased toward the default option. This additional finding implies that the Top/Flop betting mechanism might have been effective in deterring students with low motivation who started to shirk on the given task and decided to quit eventually. In complementation to our analysis on subject's valuation in the BDM-task, we run an additional binomial test excluding incomplete responses (Appendix 6), leaving our conclusion of accepting H4a unchanged.

Furthermore, we check the robustness of our analysis on response time by running the pooled OLS regression models with complete responses only, as well as fitting three generalized-OLS models with individual random effects. The results of the pooled-OLS models are robust to the exclusion of unfinished responses (Appendix 8). After running the generalized-OLS models with individual random effects, we observe few differences in coefficient size and significance between the pooled and generalized models. Table 16 shows the results of the generalized OLS-model with individual random effects.

Dependent Var: ln(response time)	**(1)** **All courses**	**(2)** **Econ Courses**	**(3)** **Math course**
Top/Flop Treatment	1.419***	1.659***	2.168***
	(0.262)	(0.244)	(0.497)
Positive Signal	0.527***	0.764***	0.227
	(0.111)	(0.144)	(0.265)
Signal-Treatment Interaction	-0.264*	-0.488**	-0.222
	(0.153)	(0.225)	(0.424)
First evaluation question	1.278***	1.394***	0.696*
	(0.137)	(0.244)	(0.393)
Order-Treatment Interaction	-0.0635**	-0.0908	-0.227***
	(0.0276)	(0.0593)	(0.0634)
Constant	0.818***	0.762**	1.065***
	(0.247)	(0.307)	(0.282)
Observations	198	131	67
Number of ID	22	22	22
Clustered on	ID	ID	ID
Demographic Controls	YES	YES	YES
Order Controls	YES	YES	YES

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 16: generalized-OLS panel model for Top/Flop sample

We observe that, in comparison to the obtained coefficients of the pooled OLS in our robustness analysis, the coefficient size for rejecting the default is halved in model (1) of the generalized-OLS regression. Subjects thus spent on average 53% more seconds rejecting the default, instead of the pooled estimate of 100%. In reconciling our findings between the two models, we find the same directions and similar magnitudes of effects for our remaining variables of interest, leaving our conclusion regarding H2a unchanged.

5.2.2 Target Betting

In testing the robustness of our results with regard to the effect of Target incentivization on default bias, we find our conclusions to be unchanged upon the exclusion of unfinished responses (Appendix 9). The robustness analysis for Target incentivization does not change our conclusion to H1b. Hence, Target incentivization and a Binary Lottery yield similar data quality in the presence of a default bias. Our supplemental analysis on the probabilistic sophistication assumption in the Target treatment confirms our conclusion of accepting H4a (Appendix 10). We proceed by investigating the robustness of our regression analysis on response time. In the same procedure as for Top/Flop betting, we estimate the pooled OLS regression model without incomplete responses and adding three generalized OLS models with individual random effects. We find some variation in terms of coefficient size in our pooled OLS models (Appendix 12). However, the direction of coefficients is not affected, leading us to conclude our accepting of H4a is robust to the exclusion of incomplete responses. Further, we estimate three generalized OLS models with individual random effects and period controls after omitting incomplete responses (Table 17).

	(1)	**(2)**	**(3)**
Dependent Var: ln(response time)	**All courses**	**Econ Courses**	**Math course**
Target Treatment	1.163***	1.154***	1.089**
	(0.219)	(0.283)	(0.440)
Positive Signal	0.654***	0.534***	0.886***
	(0.169)	(0.198)	(0.301)
Signal-Treatment Interaction	-0.323	-0.175	-0.722*
	(0.200)	(0.246)	(0.399)
First evaluation question	1.505***	1.525***	1.659***
	(0.275)	(0.220)	(0.577)
Order-Treatment Interaction	-0.0423	-0.0648*	0.0592
	(0.0325)	(0.0356)	(0.0690)
Constant	0.650*	0.754***	0.522
	(0.353)	(0.292)	(0.597)
Observations	315	210	105
Number of ID	35	35	35
Clustered on	ID	ID	ID
Demographic Controls	YES	YES	YES
Order Controls	YES	YES	YES

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 17: generalized-OLS panel model for Target sample

We obtain smaller coefficients. Our variables of interest Treatment and Signal are reduced in coefficient size but remain highly significant at the 1% level. According to model (1) of the GLS regression, subjects in the Target treatment spent on average 116% more seconds on questions when accepting the default, and 150% more seconds when issuing a positive signal. The cost of deviating from the default option based on model (1) is still substantial, as participants spent 65% more seconds relative to sticking with the default. Model 1-3 also estimate larger coefficients for the first question item, indicating that subjects spent more time pondering their answer on the first item presented to them. We conclude that our conclusion to accept H2b remains unchanged as a result of our robustness analysis.

6 Discussion

Under the implementation of a one-directional default bias, we find TFTB incentivization to enhance data quality with regards to participants effort measured in response time for both types of bets. Additionally, we find encouraging evidence that the probabilistic sophistication assumption holds in practice. We need to acknowledge that the incentives were very weak, subjects knew only 3 rewards could be obtained in total. Therefore, the low probability of winning the reward might have affected subjects' motivations significantly. Considering that subjects in the control groups had flat-fee incentivization, they might have felt less uncertainty in their perception of obtaining the reward. For subjects in the Top/Flop treatment, uncertainty was somewhat higher as they knew their probability of being rewarded depended on their betting performance. This and the weak incentives might have had adverse effects to subject's motivation resulting in higher attrition relative to the control group. Further research might provide certain incentives in the sense of piece-rate pay to ensure subjects are sufficiently incentivized. In the incentive literature, it is known that one should rather provide sufficient incentives or abstain from an implementation of incentives (Gneezy & Rusticini 2000).

An additional limitation regarding our results is that the default bias did not produce enough differences in signals to shed light on the effects of TFTB incentivization on truth-telling. Due to the sole implementation of a one directional default-bias, we could not detect whether the opposite direction of default bias would have changed our results. Since a-priori sample size concerns required trade-offs in terms of the number of treatment and control groups, we stuck with a naysayer default bias. Access to the historical data would have indicated for which direction the default bias might have been more effective, but as mentioned above we were unable to gain access to the course evaluation data at the ESE. Therefore, future research using the implementation of a default bias should ensure a combination of all possible biases to eliminate the uncertainty related to a one-directional default bias. Another reason for the relatively stronger default bias in the treatments could be unintentional experimenter demand effects of the default option. Given that subjects were unfamiliar with the method, the default option could have been interpreted as a suggestion by the experimenter.

We experienced difficulties in recruiting participants, to obtain the current sample we notified potential subjects on 3 different occasions. Our first resort was contacting the Professors teaching the course asking them to post the survey with a pre-written text on ESE's online learning platform. All Professors followed suit except for one. Additionally, we posted the survey link and text in each cohorts Facebook group twice within 5 days to increase the number of participants. This provides a further indication that the incentives might have been too low recruit a large group of students[3] to participate in our survey. Furthermore, significantly more year 2 students participated in the survey. We think this could be due to two reasons. First, year two students could have more routine in their student life, leaving them more available to participate in a short survey at the end of the academic year. Second, year 2 students might have been more willing to participate, because during their second year they also conducted assignments using surveys. Hence, lowering their incentive threshold, knowing the difficulty to collect data. Generally, we would recommend testing the TFTB method at much larger scale using Amazons Mturk or prolific. Of course, this won't allow testing the effect of the TFTB method on SETs. Further testing of the effects of TFTB on SETs could be achieved in cooperation with the Erasmus School of Economics. Surely some reservations would apply because the data is important to ESEs decision making, but we would hypothesize that an incentive such as University merchandise or vouchers redeemable on campus could be credible incentives if they came from the University directly. Further, ESE's course evaluations could be supplemented with one additional question, a Top/Flop or Target bet at the end of SETs. This would allow for a within-subject comparison of the two elicitation methods (Likert and binary Top/Flop) and students consistency in signalling their opinion.

7 Conclusion

We tested the effects of TFTB incentives on data quality in relation to a binary lottery incentive. All things considered, we find some evidence that the TFTB method can increase data quality. We found participants to spend significantly more time on questions in both TFTB treatments. Further, we are encouraged that the assumptions of the TFTB method hold in practice, specifically we found strong evidence for the probabilistic sophistication assumption. On the contrary, we did not find the hypothesized effects for the default bias and drop-out rates. Above all, we believe our results encourage further testing of the TFTB method, especially at larger scale with less restrictive sampling populations.

[3] The total number of students in both years and cohorts should exceed 1400 students.

8 References

Baillon, A. (2015). Subjective truths. In *Inaugural Addresses Research in Management Series* (pp. 1–36). Erasmus Research Institute of Management. https://doi.org/10.2307/3972508

Baillon, A. (2017). Bayesian markets to elicit private information. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1703486114

Baillon, A., Bleichrodt, H., & Granic, G. D. (2018). *Truth-telling incentives help reduce biases in survey*.

Baillon, A., & Xu, Y. (2019). *Simple bets to elicit private signals*.

Barrage, L., & Lee, M. S. (2010). A penny for your thoughts: Inducing truth-telling in stated preference elicitation. *Economics Letters*. https://doi.org/10.1016/j.econlet.2009.11.006

Beatty, P., Herrmann, D., Puskar, C., & Kerwin, J. (2007). "Don't Know" Responses in Surveys: Is What I Know What You Want to Know and Do I Want You to Know It? *Memory*, *6*(4), 407–426. https://doi.org/10.1080/741942605

Becker, G. M., DeGroot, M. H., & Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral Science*, (1), 226–232. https://doi.org/10.1002/bs.3830090304

Bertrand, M., & Mullainathan, S. (2001). Do People Mean What They Say? Implications for Subjective Survey Data. *Economics & Social Behaviour*, 67–72. https://doi.org/10.2139/ssrn.260131

Camerer, C. F., & Hogarth, R. M. (1999). The Effects of Financial Incentives in Experiments- A Review and Capital-Labor-Production Framework. *Journal of Risk and Uncertainty*, *19*(1), 7–42.

Charness, G., Gneezy, U., & Henderson, A. (2018). Experimental methods: Measuring effort in economics experiments. *Journal of Economic Behavior and Organization*, *149*, 74–87. https://doi.org/10.1016/j.jebo.2018.02.024

Coase, R. H. (1978). Economics and Contiguous Disciplines. *The Journal of Legal Studies*, *7*(2), 201–211.

Deci, E. L. (1971). Effects of externally mediated rewards on intrinsic motivation. *Journal of Personality and Social Psychology*, *18*(1), 105–115. https://doi.org/10.1037/h0030644

Falk, A., & Zimmermann, F. (2013). A taste for consistency and survey response behavior. *CESifo Economic Studies*, *59*(1), 181–193. https://doi.org/10.1093/cesifo/ifs039

Gneezy, U., & Rustichini, A. (2000). Pay enough or don't pay at all* uri. *The Quarterly Journal of Economics*, *3*(115), 791–810.

Johansson-Stenman, O., & Svedsäter, H. (2008). Measuring Hypothetical Bias in Choice Experiments: The Importance of Cognitive Consistency. *The B.E. Journal of Economic Analysis & Policy*, *8*(1). Retrieved from http://www.bepress.com/bejeap/vol8/iss1/art41

Johnson, W. T., & Delamater, J. D. (1976). Response Effects in Sex Surveys. *Public Opinion Quarterly*, *40*(2), 165. https://doi.org/10.1086/268285

Krosnick, J., & Alwin, D. F. (1987). An Evaluation of a cognitive theory of Response-Order effects in survey measurement. *Public Opinion*, *51*(2), 201–219.

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: A literature review. *Quality and Quantity*, *47*(4), 2025–2047. https://doi.org/10.1007/s11135-011-9640-9

Laguilles, J. S., Williams, E. A., & Saunders, D. B. (2011). Can Lottery Incentives Boost Web Survey Response Rates? Findings from Four Experiments. *Research in Higher Education*, *52*(5), 537–553. https://doi.org/10.1007/s11162-010-9203-2

Lazear, E. P. (1986). Salaries and Piece Rates. *The Journal of Business*, *59*(3), 405–431. Retrieved from https://www.jstor.org/stable/2352711

Machina, B. Y. M. J., & Schmeidler, D. (1992). A More Robust Definition of Subjective Probability. *Econometrica*, *60*(4), 745–780.

Marlin, J. W. (1987). Student Perception of End-of-Course. *The Journal of Higher Education*, *58*(6), 704–716. Retrieved from https://www.jstor.org/stable/1981105

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential baises, and utility. *Journal of Educational Psychology*, *76*(5), 707–754. https://doi.org/10.1037/0022-0663.76.5.707 T4

Miller, N., Resnick, P., & Zeckhauser, R. (2005). Eliciting Informative Feedback: The Peer-Prediction Method. *Management Science*, *51*(9), 1359–1373. https://doi.org/10.1287/mnsc.1050.0379

Porter, S. R., & Whitcomb, M. E. (2003). The Impact of Lottery Incentives on student survey response rates. *Research in Higher Education*, *44*(4), 389–407.

Prelec, D. (2004). A Bayesian Truth Serum for Subjective Data. *Science*, *306*, 462–467. https://doi.org/10.1126/science.1102081

Radanovic, G., & Faltings, B. (2013). A Robust Bayesian Truth Serum for Non-Binary Signals. *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, 833–839. Retrieved from http://www.aaai.org/ocs/index.php/AAAI/AAAI13/paper/viewFile/6451/7276

Rubinstein, A. (2007). Instinctive and cognitive reasoning: A study of response times. *Economic Journal*. https://doi.org/10.1111/j.1468-0297.2007.02081.x

Shevlin, M., Banyard, P., Davies, M., & Griffiths, M. (2000). The validity of student evaluation of teaching in higher education: Love me, love my lectures? *Assessment and Evaluation in Higher Education*, *25*(4), 397–405. https://doi.org/10.1080/713611436

Simonson, I., & Tversky, A. (1992). Choice in Context: Tradeoff Contrast and Extremeness Aversion. *Journal of Marketing Research*, *29*(3), 281. https://doi.org/10.2307/3172740

Smith, V. L., & Walker, J. M. (1993). Monetary Rewards and Decision Cost in Experimental Economics. *Economic Inquiry*, *31*(2), 245–261. https://doi.org/10.1111/j.1465-7295.1993.tb00881.x

Spiliopoulos, L., & Ortmann, A. (2018). The BCD of response time analysis in experimental economics. *Experimental Economics*, *21*(2), 383–433. https://doi.org/10.1007/s10683-017-9528-1

Stigler, G. J. (1984). Economics: The Imperial Science. *The Scandinavian Journal of Economics*, *86*(3), 301–313. Retrieved from https://www.jstor.org/stable/3439864

Suziedelyte, A., & Johar, M. (2013). Can you trust survey responses? Evidence using objective health measures. *Economics Letters*, *121*(2), 163–166. https://doi.org/10.1016/j.econlet.2013.07.027

Tversky, A., & Kahneman, D. (1986). Rational Choice and the Framing of Decisions. *The Journal of Business*, *59*(4), S251–S278. https://doi.org/10.1080/03057240802227486

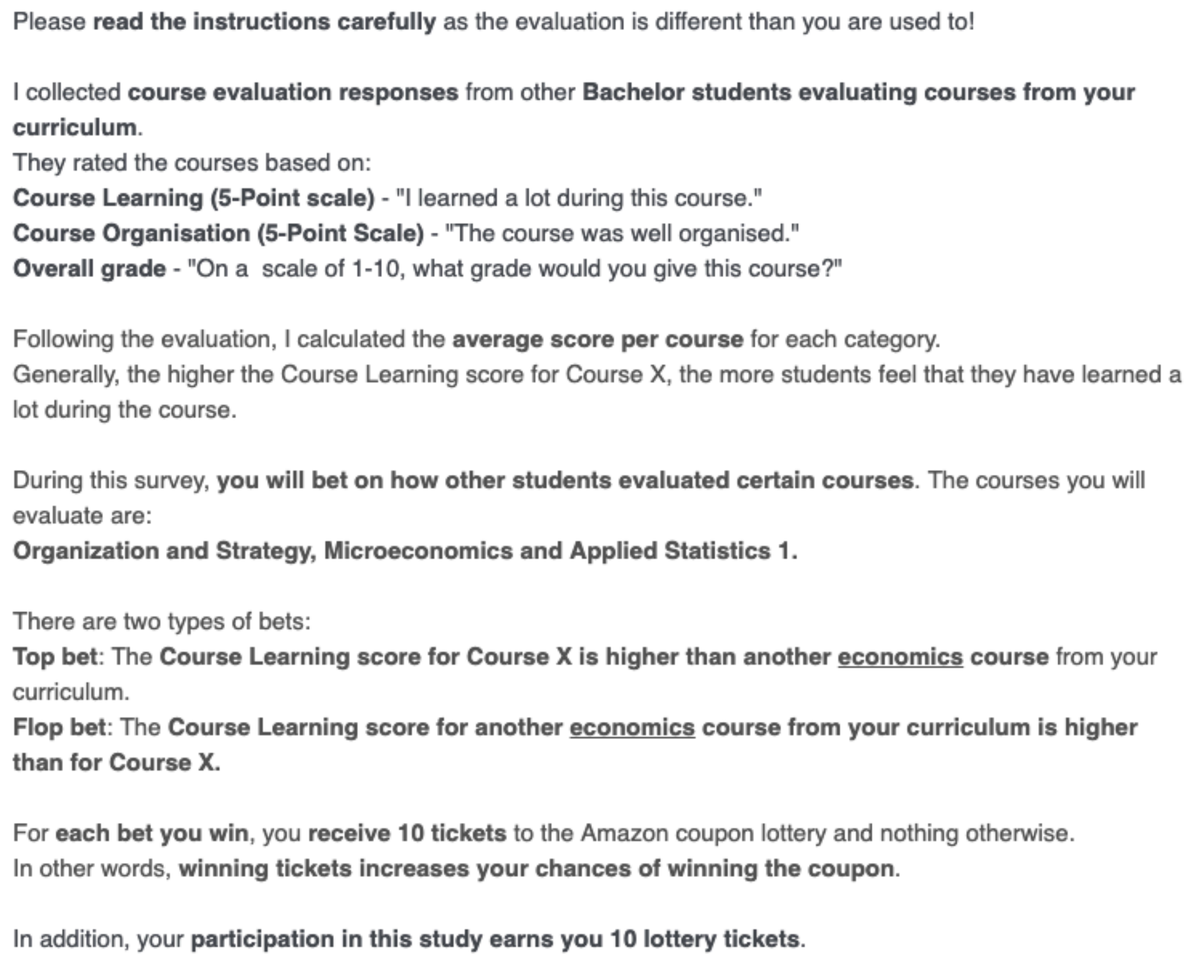
Weaver, R., & Prelec, D. (2013). Creating Truth-Telling Incentives with the Bayesian Truth Serum. *Journal of Marketing Research*, *50*(3), 289–302. https://doi.org/10.1509/jmr.09.0039

Witkowski, J., & Parkes, D. C. (2012). A Robust Bayesian Truth Serum for Small Populations. *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence.*

Xu, Y. (2016). *Will Bayesian Markets Induce Truth-telling? - An Experimental Study*. Retrieved from https://yanxu.me/wp-content/uploads/2019/06/Experimental_Test_for_BM.pdf

Appendix

Appendix 1: Instructions for Top/Flop Treatment



Please **read the instructions carefully** as the evaluation is different than you are used to!

I collected **course evaluation responses** from other **Bachelor students evaluating courses from your curriculum**.
They rated the courses based on:
Course Learning (5-Point scale) - "I learned a lot during this course."
Course Organisation (5-Point Scale) - "The course was well organised."
Overall grade - "On a scale of 1-10, what grade would you give this course?"

Following the evaluation, I calculated the **average score per course** for each category.
Generally, the higher the Course Learning score for Course X, the more students feel that they have learned a lot during the course.

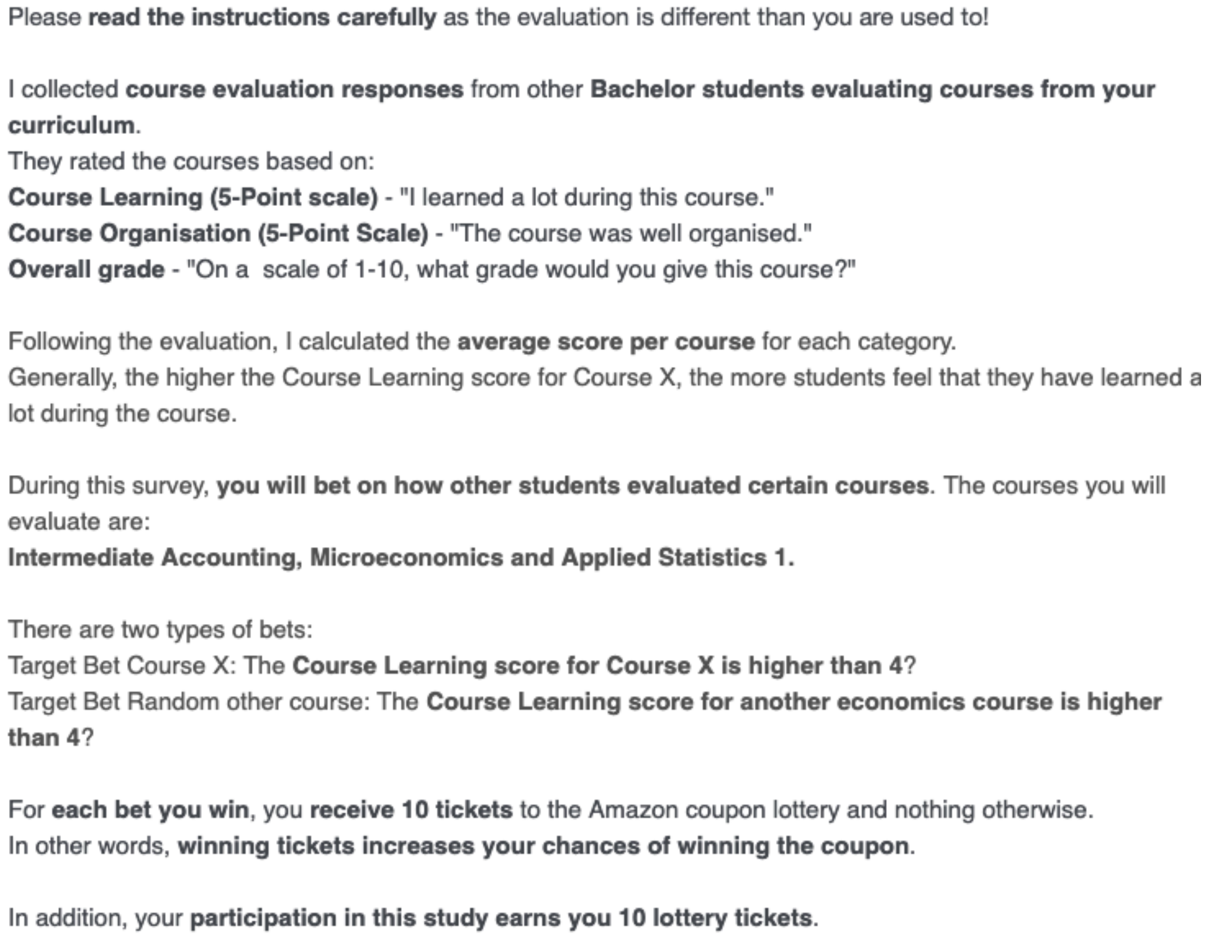
During this survey, **you will bet on how other students evaluated certain courses**. The courses you will evaluate are:
Organization and Strategy, Microeconomics and Applied Statistics 1.

There are two types of bets:
Top bet: The **Course Learning score for Course X is higher than another economics course** from your curriculum.
Flop bet: The **Course Learning score for another economics course from your curriculum is higher than for Course X.**

For **each bet you win**, you **receive 10 tickets** to the Amazon coupon lottery and nothing otherwise.
In other words, **winning tickets increases your chances of winning the coupon**.

In addition, your **participation in this study earns you 10 lottery tickets**.

Appendix 2: Instructions for Target Treatment



Please **read the instructions carefully** as the evaluation is different than you are used to!

I collected **course evaluation responses** from other **Bachelor students evaluating courses from your curriculum**.
They rated the courses based on:
Course Learning (5-Point scale) - "I learned a lot during this course."
Course Organisation (5-Point Scale) - "The course was well organised."
Overall grade - "On a scale of 1-10, what grade would you give this course?"

Following the evaluation, I calculated the **average score per course** for each category.
Generally, the higher the Course Learning score for Course X, the more students feel that they have learned a lot during the course.

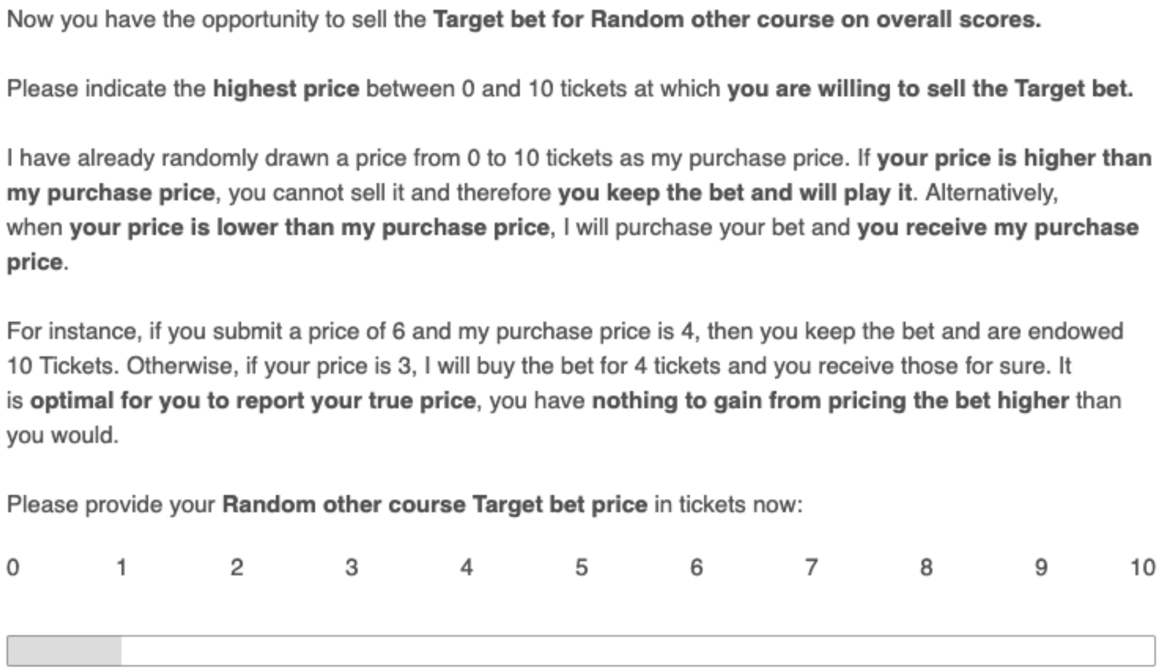
During this survey, **you will bet on how other students evaluated certain courses**. The courses you will evaluate are:
Intermediate Accounting, Microeconomics and Applied Statistics 1.

There are two types of bets:
Target Bet Course X: The **Course Learning score for Course X is higher than 4**?
Target Bet Random other course: The **Course Learning score for another economics course is higher than 4**?

For **each bet you win**, you **receive 10 tickets** to the Amazon coupon lottery and nothing otherwise.
In other words, **winning tickets increases your chances of winning the coupon**.

In addition, your **participation in this study earns you 10 lottery tickets**.

Appendix 3: Example of BDM Valuation Task in the Target Treatment



Now you have the opportunity to sell the **Target bet for Random other course on overall scores.**

Please indicate the **highest price** between 0 and 10 tickets at which **you are willing to sell the Target bet.**

I have already randomly drawn a price from 0 to 10 tickets as my purchase price. If **your price is higher than my purchase price**, you cannot sell it and therefore **you keep the bet and will play it**. Alternatively, when **your price is lower than my purchase price**, I will purchase your bet and **you receive my purchase price**.

For instance, if you submit a price of 6 and my purchase price is 4, then you keep the bet and are endowed 10 Tickets. Otherwise, if your price is 3, I will buy the bet for 4 tickets and you receive those for sure. It is **optimal for you to report your true price**, you have **nothing to gain from pricing the bet higher** than you would.

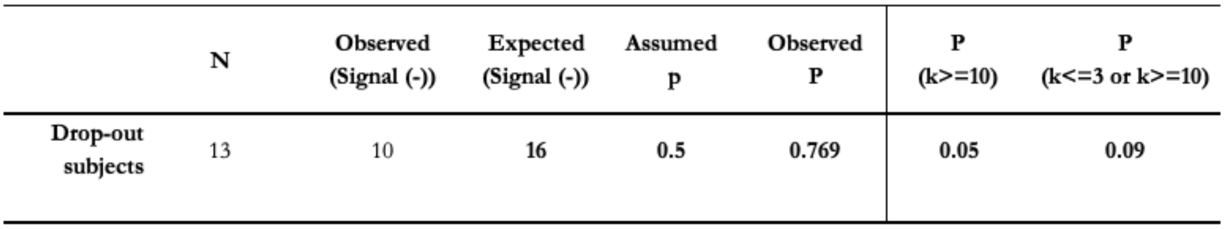
Please provide your **Random other course Target bet price** in tickets now:

0 1 2 3 4 5 6 7 8 9 10

Appendix 4: Robustness test, Fischer exact default-bias for Top/Flop sample

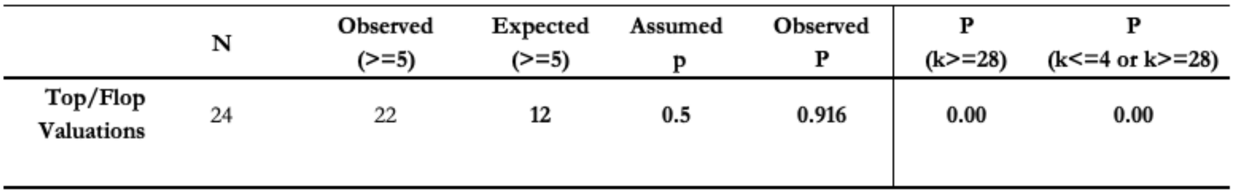
Course & Category	Treatment	Signal (+)	Signal (-)	Total	one-sided p-value
Pooled Signals					
	Control	45	39	**84**	**0.25**
	Top/Flop	22	26	**48**	
	Total	**67**	**65**		
Economics#1					
Learnings	Control	5	9	**14**	**0.49**
	Top/Flop	2	6	**8**	
	Total	**7**	**15**		
Organisation	Control	8	6	**14**	**0.546**
	Top/Flop	4	4	**8**	
	Total	**12**	**10**		
Economics#2					
Learnings	Control	9	5	**14**	**0.642**
	Top/Flop	5	3	**8**	
	Total	**14**	**8**		
Organisation	Control	7	7	**14**	**0.095**
	Top/Flop	1	7	**8**	
	Total	**8**	**14**		
Mathematics #1					
Learnings	Control	7	7	**14**	**0.246**
	Top/Flop	6	2	**8**	
	Total	**13**	**9**		
Organisation	Control	9	5	**14**	**0.416**
	Top/Flop	4	4	**8**	
	Total	**13**	**9**		

Appendix 5: Binary test for Default bias in drop-out subjects for Top/Flop betting



	N	Observed (Signal (-))	Expected (Signal (-))	Assumed p	Observed P	P (k>=10)	P (k<=3 or k>=10)
Drop-out subjects	13	10	**16**	**0.5**	**0.769**	**0.05**	**0.09**

Appendix 6: Binary test for probabilistic sophistication assumption Top/Flop (Robustness)

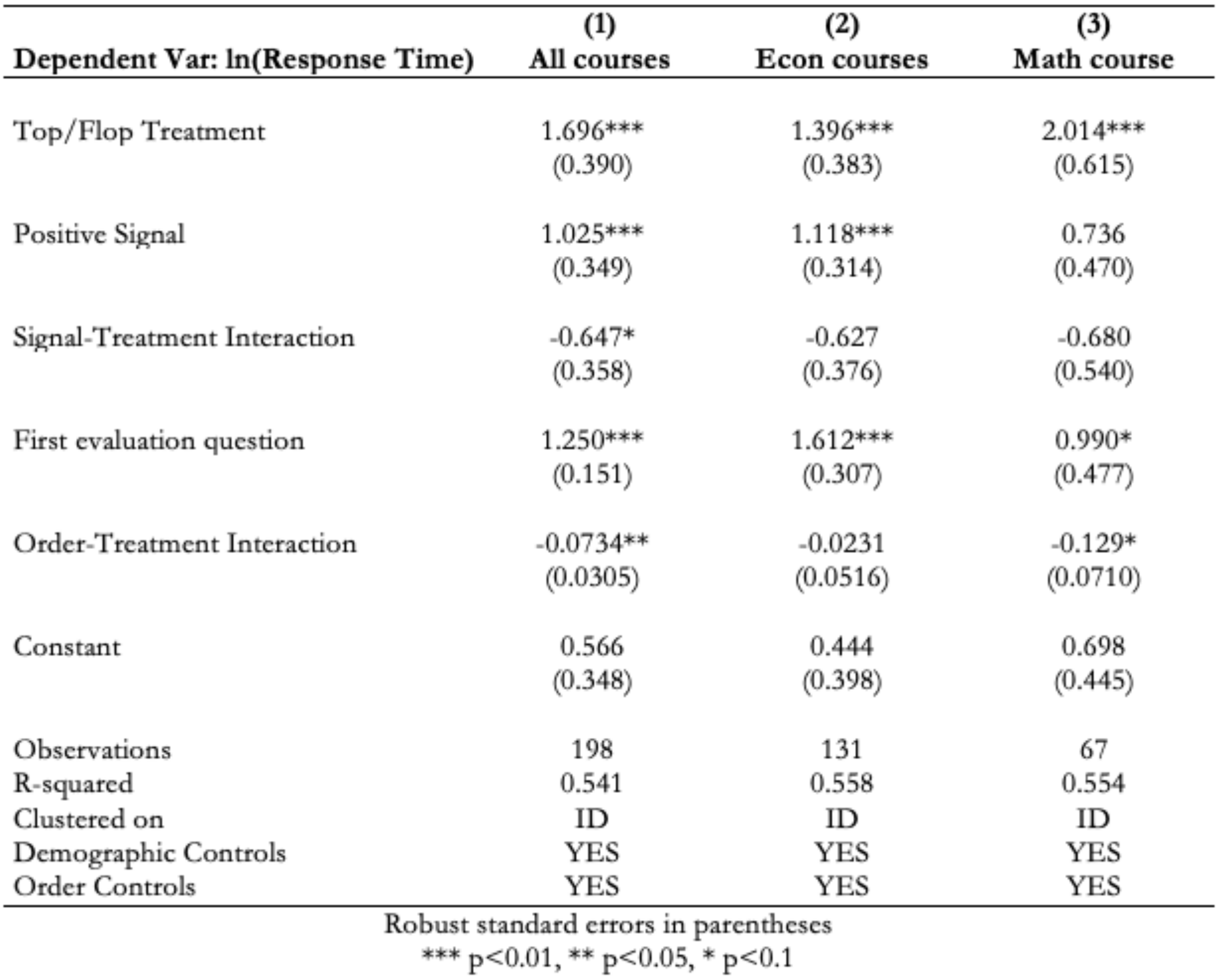


	N	Observed (>=5)	Expected (>=5)	Assumed p	Observed P	P (k>=28)	P (k<=4 or k>=28)
Top/Flop Valuations	24	22	**12**	**0.5**	**0.916**	**0.00**	**0.00**

Appendix 7: Average total response time for Top/Flop sample (Robustness)

Treatment	N	Mean (Total Response Time)	Std. Err.	Std. Dev.	CI (95%)	
Control	14	110.929	11.308	42.311	86.499	135.358
Top/Flop	8	295.25	34.742	98.266	213.098	377.402

Appendix 8: pooled OLS-regression for Top/Flop sample (Robustness)



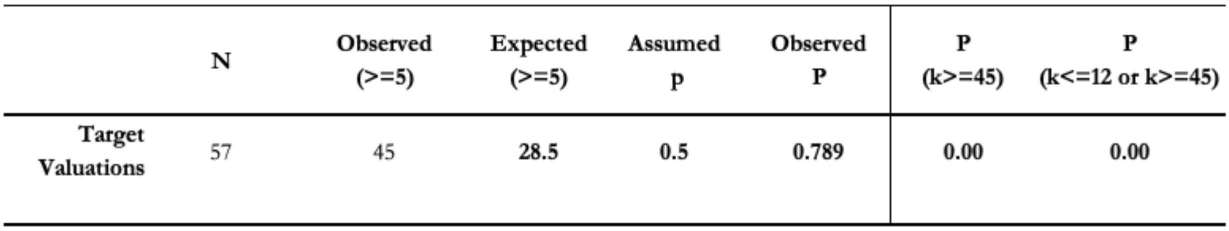
	(1)	(2)	(3)
Dependent Var: ln(Response Time)	**All courses**	**Econ courses**	**Math course**
Top/Flop Treatment	1.696***	1.396***	2.014***
	(0.390)	(0.383)	(0.615)
Positive Signal	1.025***	1.118***	0.736
	(0.349)	(0.314)	(0.470)
Signal-Treatment Interaction	-0.647*	-0.627	-0.680
	(0.358)	(0.376)	(0.540)
First evaluation question	1.250***	1.612***	0.990*
	(0.151)	(0.307)	(0.477)
Order-Treatment Interaction	-0.0734**	-0.0231	-0.129*
	(0.0305)	(0.0516)	(0.0710)
Constant	0.566	0.444	0.698
	(0.348)	(0.398)	(0.445)
Observations	198	131	67
R-squared	0.541	0.558	0.554
Clustered on	ID	ID	ID
Demographic Controls	YES	YES	YES
Order Controls	YES	YES	YES

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Appendix 9: Robustness test, Fischer exact default-bias for Target sample

Course & Category	Treatment	Signal (+)	Signal (-)	Total	one-sided p-value (Fischer exact)	Mean diff.	one-sided p-value (t-test)
Pooled Signals							
	Control	84	36	**120**	**0.006**		
	Target	47	43	**90**			**P(Z>z)**
	Total	**131**	**79**			**0.178**	**0.004**
Economics#3							
Learnings	Control	17	3	**20**	**0.000**		
	Target	3	12	**15**			**P(Z>z)**
	Total	**20**	**15**			**0.65**	**0.000**
Organisation	Control	17	3	**20**	**0.000**		
	Target	3	12	**15**			**P(Z>z)**
	Total	**20**	**15**			**0.65**	**0.0001**
Economics#4							
Learnings	Control	15	5	**20**	**0.433**		
	Target	10	5	**15**			**P(Z>z)**
	Total	**25**	**5**			**0.083**	**0.295**
Organisation	Control	15	5	**20**	**0.281**		
	Target	9	6	**15**			**P(Z>z)**
	Total	**25**	**11**			**0.15**	**0.172**
Mathematics #1							
Learnings	Control	11	9	**20**	**0.118**		
	Target	12	3	**15**			**P(Z<z)**
	Total	**23**	**12**			**-0.25**	**0.0615**
Organisation	Control	12	8	**20**	**0.324**		
	Target	11	4	**15**			**P(Z<z)**
	Total	**23**	**12**			**-0.133**	**0.2054**

Appendix 10: Binary test for probabilistic sophistication assumption Target (Robustness)

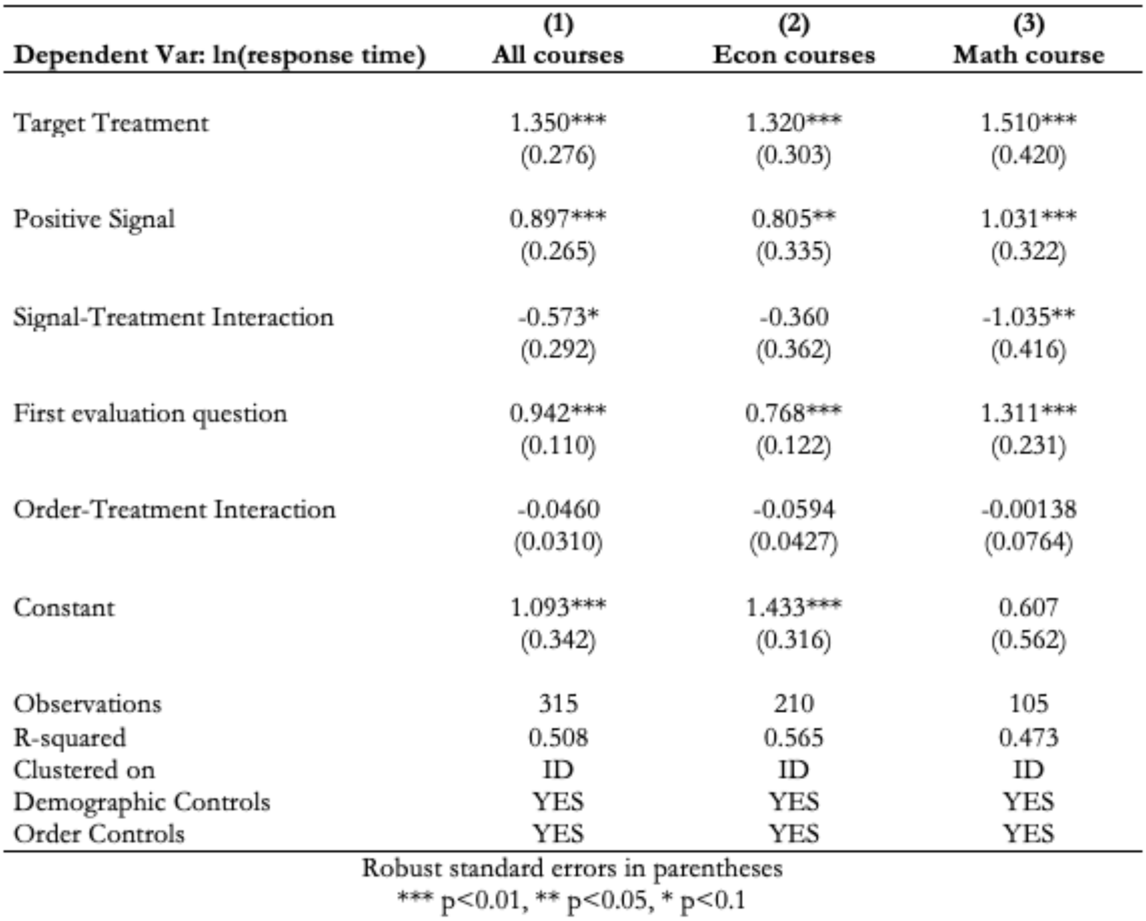


	N	Observed (>=5)	Expected (>=5)	Assumed p	Observed P	P (k>=45)	P (k<=12 or k>=45)
Target Valuations	57	45	**28.5**	**0.5**	**0.789**	**0.00**	**0.00**

Appendix 11: Average Response Time in Target sample (Robustness)

Treatment	N	Mean (Total Response Time)	Std. Err.	Std. Dev.	CI (95%)	
Control	20	126.9	11.354	50.778	103.135	150.665
Target	15	249.467	23.72	91.868	198.592	300.341

Appendix 12: pooled OLS-regression for Target sample (Robustness)



	(1)	(2)	(3)
Dependent Var: ln(response time)	**All courses**	**Econ courses**	**Math course**
Target Treatment	1.350***	1.320***	1.510***
	(0.276)	(0.303)	(0.420)
Positive Signal	0.897***	0.805**	1.031***
	(0.265)	(0.335)	(0.322)
Signal-Treatment Interaction	-0.573*	-0.360	-1.035**
	(0.292)	(0.362)	(0.416)
First evaluation question	0.942***	0.768***	1.311***
	(0.110)	(0.122)	(0.231)
Order-Treatment Interaction	-0.0460	-0.0594	-0.00138
	(0.0310)	(0.0427)	(0.0764)
Constant	1.093***	1.433***	0.607
	(0.342)	(0.316)	(0.562)
Observations	315	210	105
R-squared	0.508	0.565	0.473
Clustered on	ID	ID	ID
Demographic Controls	YES	YES	YES
Order Controls	YES	YES	YES

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1