



UNIVERSIDAD
DE GRANADA

Tesis entregada para la obtención del título de
Doctor en Ciencias Económicas y Empresariales

EXPERIMENTAL EVIDENCE ON CHOICES
RELATED TO DIETARY HABITS, ANTICIPATED
DISCRIMINATION AND CHARITABLE
CONTRIBUTIONS

por
Ángela Sánchez González

Programa de Doctorado en Ciencias Económicas y Empresariales

Editor: Universidad de Granada. Tesis Doctorales
Autor: Ángela Sánchez González
ISBN: 978-84-1195-268-2
URI: <https://hdl.handle.net/10481/90796>

The present document is organized as follows:

Summary of the Thesis

[Page 4]

1. Chapter 1: Anticipated discrimination, choices and performance.
[Page 6]
2. Chapter 2: Group identity and charitable contributions: experimental evidence.
[Page 37]
3. Chapter 3: Improving healthy eating in children: experimental evidence.
[Page 50]
4. Bibliography
[Page 88]
5. Appendix
 - Appendix Chapter 1
[Page 99]
 - Appendix Chapter 2
[Page 110]
 - Appendix Chapter 3
[Page 113]

Summary of the Thesis

This thesis studies the effect of different economic incentives on subjects' decisions within three different settings: *i)* labor markets, *ii)* charitable giving, and *iii)* dietary habits. The first two chapters involve real effort task laboratory experiments, whereas chapter three is a field experiment. The variables studied include, among others, subjects' performance, choices under potential discrimination, and children's responsiveness to a new system of food assessment. The analysis was carried out through non-parametric tests, microeconomic modelling and econometric regressions.

Chapter one studies experimentally whether potential perceived discrimination affects decisions in a labor-market setting with different stereotypes. Participants are assigned to a seven-person group and randomly allocated a role as a firm or worker. In each group, there are five workers and two firms. The only information firms have about each worker is a self-selected avatar (male, female or neutral) representing a worker's gender. Each firm then decides which worker to hire. Depending on the treatment workers are informed about the task to be performed, either math-related, emotion-related, or they will receive no information at all. Results show that female workers react to potential discrimination when they know the task is math-related, but not otherwise. Men choose similar avatar patterns regardless of the task. Men do perform at much higher levels in the math-related task, but there is no difference in performance in the emotion-recognition task, where there is a strong female stereotype.

In chapter two I look at group identity as a trigger to increase charitable giving. Although group identity has been studied under different related settings, like in-

group/out-group interactions or in Public Goods Games, little has been said on its effect on charitable behavior. With that objective in mind, I conduct a 2x2 laboratory experiment in which subjects do a real effort task to generate money with the possibility of donating part, all, or none of that money to a charity. Depending on the treatment, the real effort task is done individually or in groups of four people. I also vary whether the donation –which is made privately and individually regardless of the treatment– is either transferred as a single individual donation or added to the rest of group members' contributions and thus, transferred as a group donation. Results indicate that doing the task in groups increases the intensive and extensive margins of charitable giving for both individual and group donations. Results also show that donating as part of a group has an effect on the amount donated only when group identity was previously enhanced.

Finally, chapter three roots within the literature on children's healthy dietary habits and incentives to improve their choices. In particular, the field experiment studies the effect of non-monetary incentivize for children to make healthier food choices at school. Previous interventions have typically paid participants for their participation, but this often may not be feasible, neither sustainable in the long run. We introduce a system in which food items are graded based on their nutritional value, involving parents or classmates as change agents by providing them with information regarding the food choices of their children or friends. We find parents' involvement in the decision process to be particularly beneficial in boosting healthy food choices, with very strong results that persist months after the intervention.

1. CHAPTER 1

ANTICIPATED DISCRIMINATION, CHOICES, AND PERFORMANCE: EXPERIMENTAL EVIDENCE*

1.1 INTRODUCTION

There are strong differences across gender in the likelihood one pursues science, technology, engineering, or mathematics (STEM) careers, with males choosing courses and degrees with a strong mathematical component far more often than do females. These differences emerge at high school (Buser, Niederle and Oosterbeek, 2004; Joensen and Nielse, 2011) and remain at the college level (Zafar, 2013). This gap further widens at the graduate-school level (Hill *et al.*, 2010).

Different factors have been proposed to explain the low representation of women in science. First, gender differences in preferences (Croson and Gneezy, 2009; Niederle and Vesterlund, 2007; Niederle and Vesterlund, 2011) have been extensively analyzed in the literature as an explanation for the self-selection into math-oriented activities. Second, it is possible that men have better aptitudes for science-related tasks; however, the evidence is not conclusive on this matter. Some studies find that males perform better, particularly with respect to being in the top of the distribution than females (Hedges and Nowel, 1995; Xie and Shauman, 2003; Ellison and Swanson, 2010; Guiso, Monte, Sapienza, and Zingales, 2010; Grosse, Riener, and Dertwinkle-Kalt, 2014), while other studies indicating that there are no significant gender differences in performance (Hyde and Mertz, 2009; Hyde *et al.*, 2008; Guiso *et al.*, 2008). According to Bordalo, Coffman, Gennaioli, and Shleifer (2016), men being over-represented at the

highest performance levels leads a stereotypical thinker to exaggerate the magnitude of mean differences. This could potentially lead to the stereotype that male performance in math is high, while female performance is poor.

Stereotypes (Bordalo *et al.*, 2016; Coffman, 2014) and related discrimination issues (Bohnet, van Geen, and Bazerman, 2016; Reuben, Sapienza, and Zingales, 2014) could also play a major role in subjects' decisions since they could affect beliefs about future earnings. A recent and growing literature points to expectations as an important predictor of educational choices, showing that, along with ability perceptions, beliefs about future earnings are one of the determinants of major choices (Arcidiacono *et al.*, 2012; Zafar; Wiswall and Zafar, 2015).

While stereotypes and labor-market discrimination have been already extensively analyzed in the literature, nevertheless little is known regarding whether and how people react when they anticipate potential discrimination. This anticipation could be even more important than the discrimination itself for the agents' decision process. Consider, for example, a woman who, when planning what major to choose, anticipates that she will face discrimination (either a lower probability of being hired or getting a lower salary than her male counterpart) in a particular field. Hence, reacting strategically to this perceived discrimination could reduce the likelihood of choosing a degree that relates to some particular profession.

The goal of this paper is to study experimentally whether men and women react by making strategic decisions when they face a situation of potential discrimination. In particular, we study whether subjects in a labor-market setting hide (or reveal) their gender, perhaps in anticipation of a lower probability of being hired. Moreover, we also study how the decision to hide one's gender depends

on the task. We take this as a starting point of our experimental design, where participants were allocated to a group of seven people and randomly assigned a role as a firm or worker. In each group, there were five workers and two firms. Firms could hire one worker to perform a task in a subsequent stage. The only information that firms had about the workers was an avatar signaling a worker's gender. Prior to the firms making their decisions, the workers chose the avatar that would represent them in the market. Each worker could choose: (i) a male avatar, (ii) a female avatar, or (iii) a neutral avatar. Workers' profits were larger if they were hired by a firm than if they were not, and this was common information to all workers.

In our design, we vary the information about the task that workers receive when they pick the avatar. In the first treatment, workers know that they will be hired to perform a mathematical task that consists of adding five two-digit numbers over a five-minute period. The task is the same in the second treatment, but the workers did not receive information about the task. In the third treatment, workers know that they will be performing a task consisting of identifying emotions depicted on individuals' faces, which carries a female gender stereotype (Gigerenzer *et al.*, 2013; Halladay, 2017; Bordalo *et al.*, 2016a).

Our study makes two principal contributions. First, we explore whether a potential perception of discrimination can affect the gender of the avatar chosen by people who wished to be hired. Given the stereotype that males are better at math-related tasks and that females are better at emotion recognition, one might expect that male avatars would be the typical choice in the summing-numbers task and that female avatars would be the typical choice in the emotion-recognition task. Second, we demonstrate very clear patterns regarding the

performance on these tasks across gender. Additionally, we also report hiring patterns for different distributions of avatars available for hiring.

Our results show that men and women react differently when they face potential discrimination. When the task is math-oriented and workers have information regarding the task, males do indeed self-identify almost twice as often as females. In fact, male avatars were chosen 54.0% of the time and female avatars were chosen 23.0% of the time. However, these gender differences disappear in the other two treatments and people largely choose same-gender avatars. When the workers do not know the task prior to the avatar choice, the proportions of male and female avatars chosen were identical, at 42.9%. Finally, one might expect males to pretend to be females in the emotion-recognition task, but they do not do so. Although females chose the female avatar in most of the cases (64.1%), the majority of males (67.3%) still picked the male avatar. So, the proportion of male avatars (46.7%) is actually a bit *higher* than the proportion of female avatars (39.0%), despite the fact that women are considered to be better at identifying emotions.¹ These results would suggest that most women choose avatars strategically, hiding their gender when the task is male-oriented, but that men fail to do so in the emotion-recognition task. We discuss possible reasons for this in the results section.

Second, we observe interesting results with respect to performance on our tasks. We find a dramatic difference across gender in the adding-numbers task; males (females) average 9.06 (6.00) successful sums. This is indeed consistent with the stereotype and shows an *ex-post* failure by firms to maximize income. In great contrast, we find virtually no difference in mean performance in the

¹ Our own survey evidence, presented in Section 5, strongly indicates that the emotions task carries a clear female stereotype.

emotions-recognition task, with males (females) averaging 12.00 (12.01) successes. This is certainly not consistent with the clear prevailing stereotype.

Third, even considering the limitation of our results regarding the hiring decisions, we see no evidence of discrimination against female avatars. Perhaps surprisingly, firms do not discriminate against female avatars in the math task, despite (female) workers' expectations. There is not gender discrimination in the emotions-recognition task either.

The remainder of the paper is organized as follows. In Section 2 we explain the experimental design. Section 3 states the behavioral predictions. Section 4 shows the main results and Section 5 provides some discussion. We conclude in Section 6.

1.2 EXPERIMENTAL DESIGN AND PROCEDURES

1.2.1 Experimental design

The experimental design consists of three treatments: the Math-Information Treatment (*MIT*, hereafter), the Math-No-Information Treatment (*MNIT*, hereafter), and the Emotions-Information Treatment (*EIT*, hereafter). All three treatments involved a real-effort task.

In *MIT*, participants were allocated into groups of seven people and randomly assigned a role as a firm or a worker. Each group was composed of five workers and two firms. Each firm could hire only one of the five workers, who would then add five two-digit numbers for five minutes (Niederle and Vesterlund, 2007). Participants could not use calculators or scratch paper. Both firms and workers were paid on a piece-rate basis. Firms and workers who were hired were paid 1.5 GBP for each correct sum that workers did. Workers who were *not* hired by a firm would do the same task but would only be paid 0.5 GBP per correct

sum. We chose adding five two-digit numbers because this task is related to an area for which there is the stereotype that men perform better (Correll, 2001; Rudman *et al.*, 2001; Kiefer and Sekaquaptewa, 2007).

When deciding whom to hire to perform the task, the only information firms had about a worker was an avatar signaling the worker's gender.² Each worker had previously chosen an avatar (male, female, or neutral) that would represent him or her in the market.³ We chose to use avatars as a key feature of our design since we considered that it would provide a good environment for workers to go to the labor market represented by their preferred gender without having to state it explicitly.^{4,5}

When choosing the avatar, workers would have information regarding: (i) the task they would perform, (ii) the size and composition of the group (number of firms and workers), (iii) the fact that each firm could only hire one worker, (iv) the payoff scheme and (v) the fact that firms did not know that the avatar was chosen by the worker. Figure 1 provides a summary of all stages of the experiment.

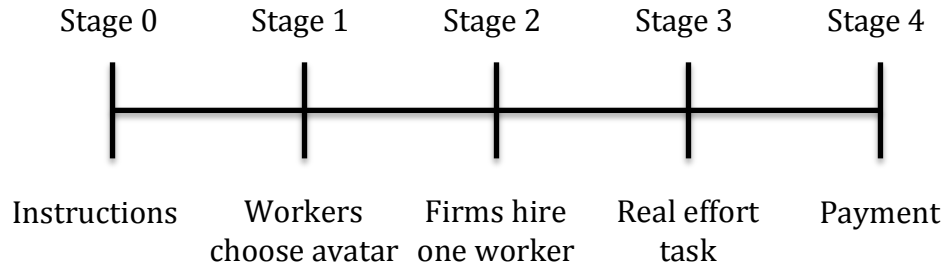
² Workers in the experiment were not explicitly told that the avatar was signaling their gender. They were just told that there would be an avatar that would represent them in the market. See the instructions in Appendix A for more details.

³ See Appendix B for the avatar alternatives.

⁴ Lim and Harrell (2015) investigated the behavioral patterns of individuals who construct their avatars, including preferences for a particular avatar gender and looked at the connections of these behavioral patterns with the participants' true gender identities. In our design, preferences would be the same across treatments, changing only the strategic considerations of the avatar.

⁵ The reason behind offering the possibility of a neutral avatar was to present an intermediate mechanism for a worker to not reveal his or her gender without active misrepresentation. Identity considerations (Akerlof and Kranton, 2000; Charness, Rigotti, and Rustichini, 2007; Chen and Li, 2009) suggest that it is likely to be costly for one to choose an avatar that is not one's actual gender. We felt that while choosing a neutral avatar was also somewhat costly in relation to one's identity, this cost would be considerably lower than choosing an avatar from the other gender.

Figure 1. Structure of the experiment.



MNIT was the same as *MIT* except that workers did not know, when choosing the avatar, the nature of the task they had to perform in the third stage.⁶ The information workers had in this treatment was: (i) the size and composition of the group (number of firms and workers), (ii) the fact that each firm could only hire one worker, (iii) the fact that their earnings could potentially be three times larger if they were hired than if they were not and (iv) the fact that firms did not know that the avatar was chosen by the worker.

EIT was the same as *MIT* with the only difference being that the task performed by the workers in the third stage was not math related. In *EIT*, workers would do a real effort task consisting of identifying emotions from faces. The facial emotion task utilizes professionally classified images obtained from The Great Good Science Center at the University of California, Berkeley.⁷ Workers were shown 15 faces, each appearing on the screen for two seconds, and subjects attempted to correctly select the depicted emotion out of four options. Subjects had 20 seconds to submit an answer after each image was displayed. They were (accurately) told that the same emotions could repeat but the same image would never appear more than once. Similar to the two previous treatments, firms and hired workers would earn 1.5 GBP per correct emotion identified and each non-

⁶ Note that in this treatment firms did know the task that workers would perform when hiring them.

⁷ We thank Brianna Halladay for granting us access to the material she had obtained from the Center.

hired worker would earn 0.5 GBP per emotion. This task was chosen since, contrary to the math task, it carries a gender stereotype that women will perform better (Halladay, 2017; Gigerenzer *et al.*, 2013, Bordalo *et al.*, 2016a).

1.2.2 Procedures

The experiment was conducted at the University of Exeter with 434 participants, who were recruited using the online recruitment system ORSEE (Greiner, 2015). The experiment was programmed and conducted with z-Tree (Fischbacher, 2007). We conducted 20 sessions of 21 subjects (6 firms and 15 workers) each; we had one session with 14 participants (4 firms and 10 workers). Subjects could see the composition of the session when entering the lab. The gender composition of all sessions was always the same, 50% male participants and 50% female participants.⁸ No subject participated in more than one session. On average, each person received 10.5 GBP for a session that was 45 minutes or less.

1.3 BEHAVIOURAL PREDICTION

We first consider the case when the task is math related and this is common information. Under the standard assumption of selfish preferences, workers would be maximizing their expected payoffs when choosing the avatar. Given the structure of the experiment, it is always better for a worker to be hired.

Thus, in accord with previous experimental results that show that there is discrimination against women in environments in which the job is math-oriented

⁸ Note that in each session there was one extra male or female participant. We chose the gender of the extra participant in each session randomly). However, note that participants were told that they would be allocated to groups of seven subjects. So, even if they could see the gender composition of the session, it was impossible for them to know the gender composition of their group.

(Reuben *et al.*, 2014), we hypothesize that workers will expect discrimination against females; a natural prediction is that more subjects will choose the male avatar over the female one. This larger percentage of male avatars chosen might come from males insisting on claiming their true gender and females choosing to hide theirs.

Hypothesis 1: *Males will tend to reveal their gender and females will have a significant tendency to hide theirs in MIT, resulting in a greater fraction of male avatars selected than the overall proportion of males in the population.*

Regarding behavior in *MNIT*, there could be (at least) two different perceptions when workers do not know the task. One possibility is that workers do not anticipate any sort of discrimination since they do not have information about what they will do in the second stage. If this were the case, workers might be expected to report their true gender (given even a small ‘cost’ of misrepresenting the truth). A second option would be that (female) workers anticipate discrimination in the general case, even when there is uncertainty about the job. This would lead to workers in *MNIT* behaving as in *MIT*. Given the lack of information about the task, we did not expect that workers would anticipate discrimination against women when uncertainty is present; due to the potential costs related to identity issues, we would expect similar proportions of chosen male and female avatars.

Hypothesis 2: *The same fraction of males and females will reveal their true gender in MNIT. This will result in a similar percentage of male and female avatars selected.*

Next, we consider the case of *EIT*. Given the nature of the task, subjects could anticipate three possible outcomes. First, participants might expect males to always have an advantage in the labor market, even when the task is more

female-oriented. Second, workers might anticipate that there will be no discrimination in the hiring process. Finally, positive discrimination toward female workers could be expected. In this case, male workers might hide their true gender while female workers might truthfully report theirs, resulting in a larger fraction of female avatars chosen. Given the stereotype that women do better in the emotions-recognition task (Halladay, 2017; Gigerenzer *et al.*, 2013; Bordalo *et al.*, 2016a), we expect workers to anticipate positive discrimination towards females, choosing the female avatar more often than in other treatments.

Hypothesis 3: *In EIT, both males and females will anticipate positive discrimination towards women in the labor market and will choose the female avatar. This will result in a larger percentage of female avatars selected.*

Finally, we focus on the decisions of the firms. Recall that the only information firms have about the workers when they decide whom to hire is the avatar. As stated in the analysis of workers' behavior, we predict that more workers will choose a male avatar in *MIT* and a female avatar in *EIT*. Given the low proportion of female avatars in the labor market in *MIT*, there would be only a limited number of cases with comparisons across types of gendered avatars. One might therefore consider that firms would see a male avatar as being less meaningful. To the extent that this is true, firms might be less likely to discriminate in the market based on the avatar; the same reasoning applies to *EIT*. In *MNIT* we hypothesized a similar fraction of male and female avatars in the market. This would make the information from the avatar appear more meaningful, and, given the previous evidence on discrimination in a mathematical task (Reuben *et al.* 2014), we conjecture that more male avatars will be hired in this treatment.

Hypothesis 4.1: *More male avatars will be hired in MNIT.*

Hypothesis 4.2: *There will be no discrimination in MIT and EIT.*

1.4 RESULTS

This section is structured as follows. We first analyze the avatar choice. Second, we study workers' performance. Then, we look at firms' decisions regarding whom to hire. Finally, we consider worker profits in relation to avatar choices.

Table 1 summarizes the avatar choices made by the workers. Specifically, we report the percentage of workers choosing a male, a female, or a neutral avatar. We do the analysis for both male and female workers.

Table 1. Summary statistics

	<i>MIT</i>			<i>MNIT</i>			<i>EIT</i>		
	<i>Avatar</i>			<i>Avatar</i>			<i>Avata</i>		
									<i>r</i>
	Male	Fema	Neutr	Male	Fema	Neutr	Male	Fema	Neutr
		le	al		le	al		le	al
<i>Total</i>	54.00	23.00	23.00	42.86	42.86	14.29	46.67	39.05	14.28
	%	%	%	%	%	%	%	%	%
<i>Males</i>	68.00	10.00	22.00	59.62	21.15	19.23	67.30	13.46	19.23
	%	%	%	%	%	%	%	%	%
<i>Femal</i>	40.00	36.00	24.00	26.42	64.15	9.43	26.42	64.15	9.43
<i>es</i>	%	%	%	%	%	%	%	%	%

Note: There were 50 men and 50 women in the *MIT* treatment, and 52 men and 53 women in each of the *MNIT* and *EIT* treatments. *How to read this Table:* The leftmost column shows that, in the *MIT* treatment, 68% of all males chose the Male avatar, while 40% of all females chose the Male avatar. The middle column shows that, in the *MNIT* treatment, 21.15% of all males chose the Female avatar, while 64.15% of all females chose the Female avatar. The identical proportion for female avatar choice by females in *MNIT* and *EIT* is not a copy-and-paste error.

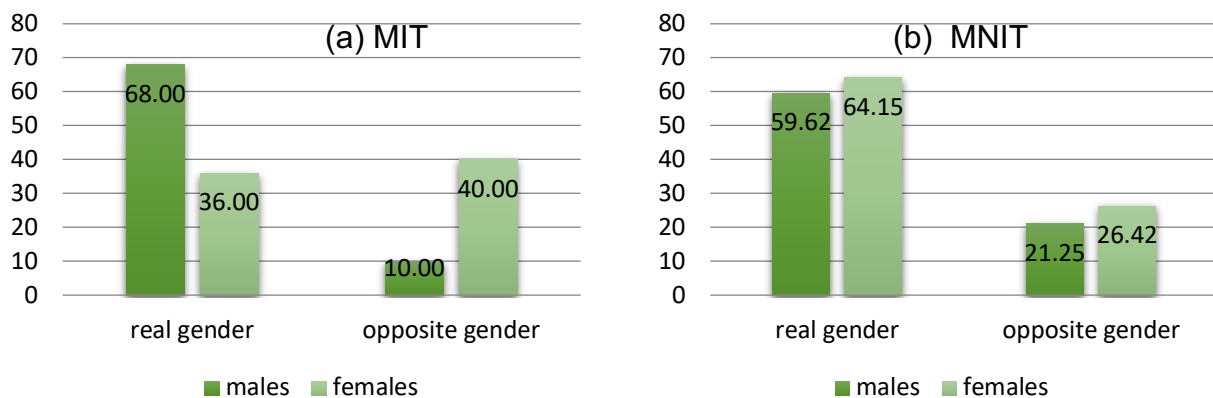
1.4.1 Avatar choices

In this section we consider the avatar choices made by each gender.

1.4.1.1 Numbers task

This subsection analyzes workers' avatar choices in the adding-numbers task according to whether the task is known or unknown. Figure 2 displays the percentage of male and female workers reporting their true and opposite gender in the numbers treatments.⁹

Figure 2: Percentage of subjects revealing gender in numbers task



Panel (a) shows that, when workers have information regarding the math-related task, males self-identify almost twice as often as females, 68.00% versus 36.00%. This difference is easily statistically significant ($Z = 3.187$, $p = 0.001$, two-tailed Mann-Whitney test).¹⁰ This largely drives the fact that the proportion of male avatars chosen in *MIT* is more than twice the proportion of female avatars, 54.00% vs. 23.00% ($Z = 4.505$, $p = 0.000$, two-sample test of proportions). These results strongly support Hypothesis 1.

Less than one-third (31.25%) of those males who hid their gender choose a female avatar, leading to only 10% of males choosing the opposite avatar, as

⁹ For an analysis of neutral avatar choices, see Appendix C

¹⁰ Here and elsewhere, we round p -values to the third decimal place; tests are two-tailed unless otherwise indicated.

shown in Panel (a). However, in the female case, twice the proportion (62.50%) of those people who hid their true gender chooses a male avatar, for a total of 40% of women choosing the opposite avatar. This difference is statistically significant ($Z = -2.022$, $p = 0.043$, Mann-Whitney test). This result seems to reflect the idea that when workers do not report their true gender, males and females behave in very different ways with females being more willing to choose the opposite avatar than males.

Result 1: *When it is common information that there is a math-oriented task, male workers are much more willing to report their true gender than are female workers.*

Matters change dramatically when there is no prior information about the math task. While the percentage of self-identifying males is not very different (59.62%; $Z = 0.876$, $p = 0.381$, Mann-Whitney test), female workers in *MNIT* choose to reveal their true gender 64.15% of the time, nearly doubling the percentage in *MIT* ($Z = -2.842$, $p = 0.004$, Mann-Whitney test). Thus, there is no significant difference in self-identification rates across gender when the task is unknown ($Z = -0.476$, $p = 0.634$, Mann-Whitney test). Furthermore, the overall proportion of male and female avatars chosen is identical (42.86% for each), supporting our Hypothesis 2.

Result 2: *When workers do not know the task, there is no significant difference between male and female workers in the percentage of people reporting their real gender and the proportion of male and female avatars chosen is the same.*

We perform regressions to better understand workers' decisions regarding the avatar in the math-related task. The first column in Table 2 reports the marginal effects from a *Probit* model in which the dependent variable is *True Gender*, a dummy taking the value 1 if the avatar that represents the true gender

of the worker was chosen and 0 otherwise. We use the following explanatory variables: *Male*, a binary covariate that equals 1 if the worker is male, and 0 otherwise; *STEM*, a dummy taking the value 1 if the worker is currently studying a degree in Science, Technology, Engineering or Mathematics, and 0 otherwise; *Treatment*, taking value 0 if *MIT* and value 1 if *MNIT*, and *Number of correct sums*. We also add an interaction between *Male* and *Treatment* to determine whether the information about the task they will perform that workers receive prior to choosing the avatar affects males and females differently; and an interaction between *Male* and *Number of correct sums* to test whether the effect of the number of correct sums differs across gender.

Table 2. Probit regression on the probability of reporting true gender in math task

	(1)	(2)	(3)
Male	0.329*** (0.094)	0.321*** (0.097)	0.169 (0.158)
STEM	-0.135 (0.088)	-0.139 (0.089)	-0.142 (0.089)
Treatment	0.281*** (0.095)	0.280*** (0.095)	0.286*** (0.095)
Male*Treatment	-0.379*** (0.125)	-0.381*** (0.125)	-0.397*** (0.124)
Number of correct sums		0.003 (0.009)	-0.011 (0.014)
Correct sums*Male			0.023 (0.018)
Observations	205	205	205
Log Likelihood	-131.458	-131.376	-130.529

Notes: Standard errors in parentheses. ***, **, and * denote significance at $p = 0.01$, 0.05 , and 0.10 , respectively.

Consistent with Result 1, the probability of picking the avatar that reveals the true gender of the worker is higher for male participants. Also, we observe that being in the treatment in which workers do not have information regarding the task increases the probability of reporting the real avatar. Finally, results

show that releasing information about the task affects male and female workers in different ways. While females report their true gender with a higher probability when the task is unknown, for males it is the opposite. This would explain the lack of gender differences in reporting the true avatar stated in Result 2, seen by the lack of significance when adding the coefficients for *Male* and *Male*Treatment* ($p = 0.610$, Wald test).

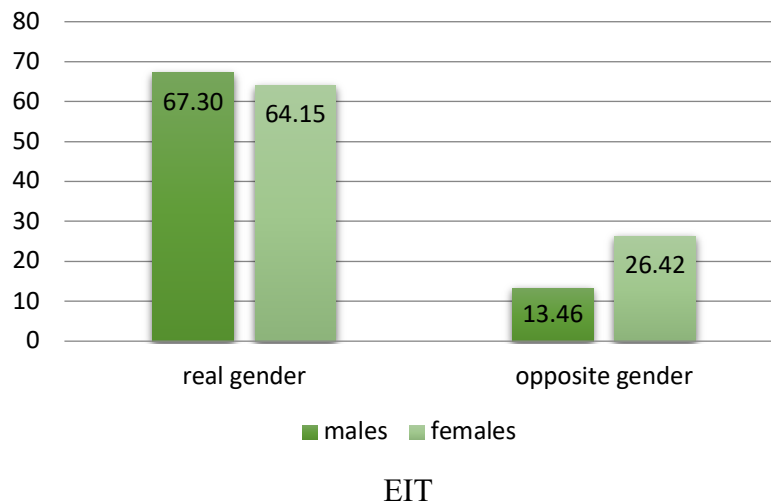
Column (2) repeats the specification in column (1), adding the number of correct sums solved by the worker in the second stage. Although the decision about the avatar was already taken when solving the sums, we used this variable as a proxy to study some potential relationship between the avatar chosen and the productivity of the worker. The coefficient of the *Number of correct sums* is not significant, indicating that productivity did not play a role in the election of the avatar. The lack of significance of productivity on the probability of self-identification remains when we introduce the interaction between the *Number of correct sums* and *Male* in Column (3). This indicates that '*posteriori*' productivity is not relevant for the avatar decision regardless of the worker's gender.

1.4.1.2 Emotion-recognition task

In this subsection, we analyze the effect on workers' avatar choices when the task is more female-oriented. As shown in Figure 3, a very similar percentage of males (67.30%) and females (64.15%) self-identify in *EIT*, leading to a similar proportion of male and female avatars chosen (46.67% and 39.05%, respectively). This does not provide support for Hypothesis 3.¹¹

¹¹ Differences are statistically significant neither for the percentage of real avatars chosen by males and females ($Z = 0.339$, $p = 0.734$, Mann-Whitney test) nor for the percentage of male and female avatars chosen ($Z = 1.115$, $p = 0.265$, test of proportions).

Figure 3: Percentage of subjects revealing gender in the emotions task



As shown in Table 1, the choice of the avatar in *EIT* is very similar to that in *MNIT*, when workers do not have information about the task. In *EIT*, 67.30%, 13.46% and 19.23% of males choose the male, female and neutral avatars, respectively; these figures are 59.62%, 21.15%, and 19.23% in *MNIT*. None of the pairwise differences across the two treatments are statistically significant ($Z = 0.814$, -1.037 , and 0.000 , $p = 0.416$, $p = 0.300$, and $p = 1.000$, respectively, test of proportions for the percentage of male, female and neutral avatars chosen).

Result 3: *When the task is female oriented, most male and female workers report their true gender and the percentage of male and female avatars chosen is very similar.*

A similar result is found for female workers' choices. In *MNIT*, the percentage of women choosing a male, female and neutral avatar is 26.42%, 64.15% and 9.43%, respectively. The equivalent percentages in *EIT* are identical, so that here there are no significant differences across treatments either ($Z = 0.000$, $p = 1.000$, test of proportions).

1.4.2 Workers' performance

This section analyzes the main determinants of workers' performance in the real-effort task. Table 3 shows the mean performance for males and females in each of our treatments.

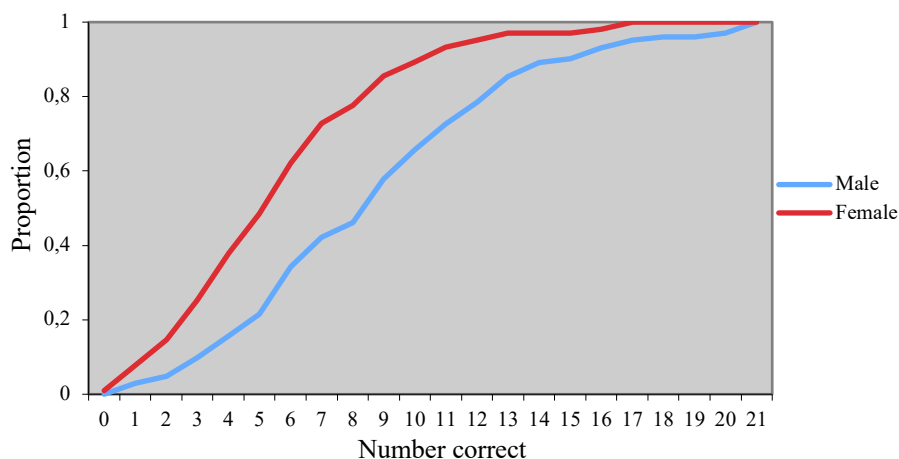
Table 3. Performance by task and gender

	<i>Numbers</i>	<i>Emotions</i>
Male	9.059 (0.453)	12.000 (0.296)
Female	6.000 (0.347)	12.019 (0.223)

Note: Standard errors in parentheses.

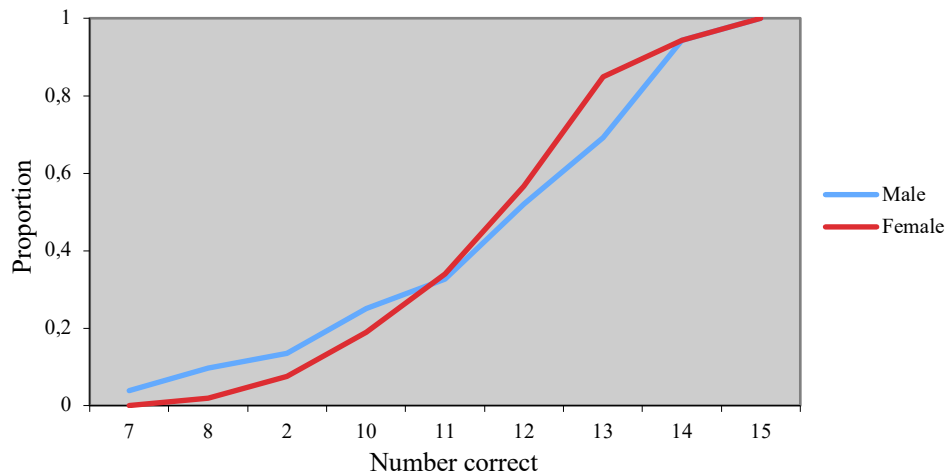
Casual inspection shows that the mean performance is quite different in the numbers task but practically identical with the emotions task. We can also explore the distribution of performance, as shown in Figures 4 and 5. In Figure 4, the Kolmogorov-Smirnov test of cumulative distributions finds a very strong difference in the addition task, $X^2_2 = 20.589$, $p = 0.000$. We observe clear first-order stochastic dominance. Consistent with some previous results, we do see a strong difference at the high end of the distribution – 15 of 102 males (14.7%) scored 14 or higher compared to 3 of 103 females (2.9%) – that is quite statistically significant ($Z = 3.152$, $p = 0.002$).

Figure 4: Cumulative distribution of performance in the math task, by gender



In Figure 5, the Kolmogorov-Smirnov test finds no serious difference in the emotions task, $X^2_2 = 2.580$, $p = 0.275$. Recall that the mean of the number correct was essentially identical (12) for males and females on the emotion-recognition task. While the differences in Figure 5 are small, they suggest that there is second-order stochastic dominance for females over males, which is consistent with the smaller variance for males shown in Table 3. Thirty-five women (34.0%) scored within one point of the mean (11, 12, or 13) compared to 23 of the men (22.5%).

Figure 5. Cumulative distribution of performance in the emotions task, by gender.



We now proceed with econometric analysis. Column (1) in Table 4 uses the data from *MIT* and *MNIT* and reports the coefficients of an OLS model in which the dependent variable is the number of correct sums solved by the worker. We include as predictors of the performance the same set of explanatory variables used in the analysis of the probability of self-identification (*Male*, *STEM* and *Treatment*). We add to the model a dummy for *Being Hired*, taking the value 1 if the worker was hired, and 0 otherwise; and *Real Avatar*, a dummy with value 1 if one reported one's true gender in the first stage of the game, and 0 otherwise. We also add the interaction between *Real Avatar* and *Being Hired*, allowing us to

assess whether subjects who were hired after hiding their gender behaved differently from those who truthfully reported it.

Individual performance is only affected by the gender of the worker, with males solving significantly more sums than females. This gender difference could be explained by the fact that, in line with previous literature (Xie and Shauman, 2003; Ellison and Swanson, 2010; Bordalo *et al.*, 2016), 83% of the workers who are performing in the top 10% of the distribution were male. Having a larger piece rate when the worker is hired does not affect her performance. The interaction between *Being Hired* and *Real Avatar* is also not significant, so it seems that workers who were hired after not self-identifying do not perform any differently in the task. This suggests that the smaller piece rate was a sufficiently strong incentive to induce workers to exert maximum effort, so that extra incentives do not affect workers' performance.

Table 4. OLS regression on performance

	(1)	(2)	(3)
Constant	5.358*** (0.714)	4.925*** (0.648)	11.917*** (0.478)
Male	3.132*** (0.583)	2.897*** (0.600)	-0.077 (0.387)
STEM	0.999 (0.705)	0.862 (0.707)	-0.083 (0.483)
Treatment	0.857 (0.580)	0.960* (0.579)	
Being Hired	-0.282 (0.01)	0.477 (0.591)	0.052 (0.657)
Real Avatar	-0.322 (0.767)		0.339 (0.513)
Being Hired*Real Avatar	1.201 (1.188)		-0.378 (0.811)
Male Avatar		0.613 (0.617)	
Observations	205	205	105
Adjusted R ²	0.119	0.123	0.043

Notes: Standard errors are in parentheses.

***, **, and * denote significance at $p = 0.01$, 0.05 , and 0.10 , respectively.

Column (2) replaces *Real Avatar* with *Male Avatar*, a binary covariate that takes value 1 if the avatar chosen by the worker was male, and 0 otherwise. Results for this second specification are the same as in Column (1), with the only significant effect coming from the *Male* dummy. Column (3) repeats the specification in column (1) using the data from *EIT* instead. In this case, the dependent variable is the number of emotions correctly identified. Results are very similar to those in the mathematical task with the only difference being that in this case the gender of the workers does not explain their performance.

Result 4: *Only the gender of the worker affects his/her performance in the math task, with males being more productive than females. The gender does not affect workers' performance in the emotions task.*

1.4.3 Hiring decisions

We now discuss the behavior of the firms. Note that, as a consequence of the endogenous choice of avatars, the distribution of candidates that firms face is different. It could happen that particular distributions affect firms' decisions when it comes to hire one worker. Given the limitation of our sample for the different distributions, our results should be seen more as initial evidence than as very reliable findings.

As a first cut we observe that, in *MIT*, the percentage of male and female avatars hired (as a proportion of all hired avatars) is 37.50 and 35.00, respectively. The equivalent figures are 38.09 and 40.47 in *MNIT*, and 38.09 and 50.00 in *EIT*.¹² This result indicates that firms hire a similar proportion of male and female avatars regardless of the treatment.¹³

¹² Differences are not statistically significant in any treatment ($Z = 0.233$, $p = 0.816$, $Z = -0.223$, $p = 0.823$, and $Z = -1.099$, $p = 0.271$ for *MIT*, *MNIT*, and *EIT*, respectively, all tests of proportions).

¹³ As we do not discuss the hiring results with neutral avatars in the text, we provide these in Appendix E.

Moreover, the proportion of male and female avatars hired by male and female firms is also not dramatically different. Pooling the numbers data from *MIT* and *MNIT*, we find that the percentage of male avatars and female avatars hired by male firms is 37.50% and 30.00%, respectively. Differences are not statistically significant ($p = 0.478$, test of proportions). The respective figures for the percentage of male and female avatars hired by female firms are 38.09% and 45.24%. Differences are not statistically significant either ($p = 0.507$, test of proportions).

There are also no dramatic differences in hiring decisions in the emotions-recognition task (*EIT*). The percentages of male and female avatars hired are 38.09% and 42.86% when the firm is male and 38.09% and 57.14% when the firm is female. Differences are not statistically significant for either case ($p = 0.753$ and $p = 0.216$, test of proportions for male and female firms, respectively). Overall, this analysis indicates that there is no discrimination in relation to female avatars in any treatment, refuting Hypothesis 4.1 and supporting Hypothesis 4.2.¹⁴

Result 5: *There is no discrimination against female avatars in either task.*

1.4.4 Avatar choices and worker profits

What profits did workers earn when choosing different avatars in each treatment? Table 5 shows workers' profits by gender and avatar.¹⁵ Males did best by choosing female avatars in the *MIT* and *EIT* treatments and by choosing male avatars in the *MNIT* treatment. The difference approaches

¹⁴ Alternative analyses on hiring decisions are reported in Appendix D.

¹⁵ Workers earn more with the emotion-recognition task than the adding-sums task, but this simply reflects the payoff calibration in the different treatments

significance only in the case of *MIT*, however (12.00 versus 6.43, $p = 0.073$).^{16,17}

Female workers also did best by choosing female avatars in the *MIT* and *EIT* treatments, but actually did best by choosing a neutral avatar in the *MNIT* treatment. The only difference that is significant is the comparison between neutral and female avatars in the *MNIT* treatment. Nevertheless, in no case was it qualitatively advisable for a female worker to choose a male avatar over a female avatar.

Table 5. Worker's profits

<i>Male workers</i>	<i>MIT</i>	<i>MNIT</i>	<i>EIT</i>
Male avatar	6.43 [34] (0.89)	9.82 [31] (1.51)	9.07 [35] (0.93)
Neutral avatar	8.23 [11] (1.71)	8.25 [10] (2.91)	9.50 [10] (1.91)
Female avatar	12.00 [5] (2.80)	7.09 [11] (1.43)	11.36 [7] (2.77)
<i>Female workers</i>	<i>MIT</i>	<i>MNIT</i>	<i>EIT</i>
Male avatar	4.82 [20] (0.96)	5.11 [14] (0.95)	12.03 [14] (1.63)
Neutral avatar	4.71 [12] (0.73)	11.20 [5] (4.35)	10.00 [5] (2.98)
Female avatar	5.97 [18] (1.20)	5.20 [34] (0.72)	12.35 [34] (1.07)

Note: Standard errors are in parentheses. The number of observations is in italics in brackets.

¹⁶ We have $p = 0.474$, and $p = 0.508$ for the male comparisons in *MNIT*, and *EIT*, respectively.

¹⁷ Note that in some cells there are only a few observations, so that the corresponding statistical tests are weak.

1.5 DISCUSSION

1.5.1 Avatar choice

First, we wished to explore whether would-be workers would react strategically to potential anticipation of gender discrimination in hiring in an adding-numbers task and an emotions-recognition task. One could argue that workers might think that the avatars were meaningless in this setting. Although they were told that firms did not know that the workers chose the avatars, they still might believe that firms believed that avatars were endogenous, since firms also were not told that the experimenter imposed the avatars on the workers. If the workers believed that firms believed that the workers did not choose their avatars, workers should mainly choose true avatars due to identity considerations (or at worst a random draw if we do not assume any identity issues). As the results show, this is not the case.

Indeed, we find that only 36% of female workers choose female avatars when they know that the task involves adding numbers, compared to 64% when they know nothing about the task. Thus, many female workers seemed to believe that choosing a female avatar would adversely affect the chance of being hired for a math task. This matches up well with the common stereotype that males are better than females at numerical or mathematical tasks. Some males did choose female avatars, but this was fairly rare (10%) when they knew the task involved adding numbers; this compares to 21% in *MNIT*.

For the emotions-recognition task, a typical stereotype is that women are better at this task and our survey results below show that this holds strongly in our sample of British students. We find that 64% of females (the same rate as in *MNIT*) chose female avatars in *EIT*. However, only 13% of male workers chose

female avatars in this stereotype situation, far less than the 40% of female workers choosing male avatars in *MIT* ($Z = 3.037$, $p = 0.002$, test of proportions).

Indeed, one might wonder why male workers don't choose female avatars in the emotions-recognition task. Perhaps it is the case that there is no female stereotype (or that the stereotype is not strong enough) for this task. To test this, we conducted an additional experiment with 40 subjects (20 males and 20 females). Similarly, to Aguiar *et al* (2009), two different boxes labeled "female" and "male" were placed at the front of a room. Each box contained 20 slips of paper. Each slip was printed with the number of correct emotions identified by the workers in the experiment.¹⁸

The decision was simply to choose the box from which to select one slip of paper. The number printed in the paper would determine subjects' payoffs. Participants were paid 0.5 GBP times the number on the slip of paper selected. This was common information. Subjects approached the boxes one by one and made their decision privately. Once the decision was made, they would show the slip of paper to the experimenter and would receive their money. The slip of paper was then put back in the corresponding box.

We find that 85% of the population took the slip of paper from the "female" box, showing that people believe that women are better than men at this task; the binomial test on the entire population gives $p = 0.000$. When we distinguish by gender, results are very similar for males and females. Eighty percent of male ($p = 0.007$) and 90% of female participants ($p = 0.000$) picked the slip of paper from the "female" box. Since this demonstrates that the emotions task was heavily

¹⁸ The 40 workers reported in the slips of paper were randomly selected from the entire subject pool that had participated in the main experiment.

considered to be a female-oriented task, we can rule out the explanation that the stereotype of this task was not prevalent.

Evidence from the sociology literature suggests that males could be reluctant to switch to a non-male avatar because male identity is associated with power and effectiveness. In Goldberg (1968), articles written by women or men were given to students for evaluation. They find that articles authored by men received higher ratings. Langford and MacKinnon (2000) find that powerful traits that are stereotyped as characteristic of men tend to be felt as good while powerless traits that are stereotyped as characteristic of women tend to be felt as bad.

Subjects in Eagly and Wood (1982) judged that men were more influential and women more easily influenced, even when given no information about the individual's roles. Eagly, Wood, and Diekmann (2000) find that there is a tendency for the specific roles occupied by men ("providers") to have more status than the roles occupied by women ("domestic"). Rashotte and Webster test the Ridgeway (1997) view that status beliefs (ideas about competence that are created by gender) arise whenever there are mixed-gender interactions and they find significantly higher general expectations for men than for women.

All in all, it may very well be true that while women have a direct stake in gender equality, men are reluctant to give up their perceived status and privileges.¹⁹ This could help to explain the male reluctance (perhaps subconscious) to switch away from the male avatar.

¹⁹ Some readers may feel that these stereotypes are outdated. Yet a recent paper (Charness and Rustichini, 2011) finds strong differences across males and females when they play the Prisoner's Dilemma in front of audiences: Men are less cooperative when in front of their peers, while women are more cooperative when in front of their peers. These experiments were conducted in California, arguably the forerunner in the women's movement.

1.5.2 Performance

While “everyone knows” that males are better at numerical and mathematical tasks, the evidence from research studies is mixed. The seminal study by Niederle and Vesterlund (2007) finds only a very small difference in performance (10.68 versus 10.15, in favor of males). Halladay (2017) finds a modest and insignificant difference ($p = 0.233$), with males doing 9.55 correct sums and females doing 8.58 correct sums. Several studies mentioned in the introduction suggest that males do perform better, particularly with respect to being at the top of the distribution; in addition, Niederle, Segal, and Vesterlund (2013) find that men perform significantly better than women in the adding task. However, other studies, including the meta-study by Hyde and Mertz (2009), find no such significant differences in general math performance. Nevertheless, to the best of our knowledge, nearly every such comparative study does find better performance by males on the adding task.²⁰

Our data indicate very strongly that there is a gender difference in performance in adding up two-digit numbers. We cannot really explain why there is such a difference, but with a significance level of $p < 0.001$ our results hardly seem accidental. The subjects in the experiment were not unusual in any way and people have not historically been suspicious of results found in a standard subject pool of British students (in this case Exeter). So we stand by our result, which is certainly consistent with the near-universal stereotype. While our results go against the prevailing orthodoxy in the literature, we do believe it is important to report non-conforming outcomes reached through standard experimental methodology.

²⁰ Even the two co-authors of the original study find a significant gender difference in performance.

On the other hand, the nearly identical performance in the emotion-recognition task is not consistent with the stereotype that women are better at this. So, in this case hiring female avatars is at least not costly in expectation to the hiring firm. Halladay (2017) finds a similar result with the identical task: Males correctly identify an average of 8.19 emotions, while females correctly identify 8.08 emotions ($p = 0.806$). So, support for this stereotype seems elusive.

1.5.3 Hiring patterns

Despite the behavioral predictions stated in Hypothesis 4.1, we observe no discrimination against female avatars.²¹ We can only conjecture regarding the forces underlying this pattern. One notion is that firms like to hire fewer common avatars. Male avatars were considerably more common in *MIT*, so this is consistent. In *MNIT*, the number of male and female avatars was the same and the hiring rates are quite similar. There were slightly more male avatars than female avatars in *EIT* and the hiring rate was higher for female avatars. In fact, relative scarcity seems to be a fairly good predictor of hiring rates. If we regress the hiring-rate data in Figure D1 in Appendix D and Appendix E (nine observations, one for each type of avatar in each treatment) against the proportion of each avatar chosen in the respective treatment, we have this simple regression:

$$\text{Hiring rate of avatar} = 60.07 - 51.21 \cdot \text{Avatar_proportion} \quad \text{Adj. } R^2 = 0.291$$

(8.94) (24.74)

²¹ Note that we should take our findings regarding discrimination with caution. The experiment was not designed with the main purpose of testing gender discrimination and has some limitations to this respect. Different firms see different distributions of workers and we do not have enough power to extract strong conclusions on this matter.

This gives $t = -2.07$ for the coefficient of the avatar proportion, significant at $p = 0.077$.²² Thus, there may be a preference for less-frequent avatars, as if firms prefer unusual workers.

What is particularly surprising is that firms made costly mistakes by hiring female avatars in the math treatments, given that male were reluctant to choose female avatars and the male workers on average completed 50% more successful summations than female workers. This is even more striking when we consider that this hiring behavior goes directly against the (in this case correct) stereotype that males perform better than females in this realm.²³

1.6 CONCLUSION

This paper considers an employment setting with excess workers, where the only information firms have about a worker is the avatar (male, female, or neutral) chosen by that worker. We examine the choice of the avatar in three treatments. The first involves adding two-digit numbers in which the workers know the task before choosing an avatar, while the second also involves adding numbers, but in this case, workers do not know the nature of the task before selecting avatars. The third treatment involves an emotion-recognition task in a setting where the workers also know the task before picking the avatar.

By looking at the avatar choices, we analyze whether workers react to a potential anticipation of discrimination and how this varies across the type of task and the information condition. Results show that males and females workers

²² An alternative regression using the hiring-rate data from the analysis in Figure 5 and Appendix C gives similar results, with $t = -1.68$ for the coefficient of the avatar proportion, significant at $p = 0.136$. The significance levels are low for the t -statistics because there are only nine observations.

²³ An alternative explanation for the lack of discrimination is that firms could conceivably be concerned about the experimenter's view of them. However, given that firms' payoffs were linked to workers' performance and that the decisions were completely anonymous, we do not think that this was a very strong driver of firms' decisions.

react differently when they face a situation of potential discrimination. We find that when the task was math-oriented (carrying a male stereotype), most women indeed decided not to reveal their true gender; presumably because they anticipated this would lead to a lower probability of being hired. However, when the task involved emotion recognition, women were much more willing to select an own avatar and did so at the same rate as men. Women also typically chose female avatars when the task was unknown. The difference in *MIT* suggests that women behaved strategically, and they chose carefully whether to reveal their real gender, depending on the situation.

However, in the emotion task, which carries a female stereotype, men were quite prone to choose a male avatar regardless. It could be that males were simply not strategic, but our sense is that this is tied to a reluctance to deny the male identity that is considered to be so useful in society. This would mean that a strong sense of power- or privilege-related gender identity (Akerlof and Kranton, 2000) inhibits males from choosing a non-male avatar, as discussed earlier. Other possible explanations include overconfidence (Lichtenstein, Fischhoff, and Phillips, 1982; Lundeberg, Fox, and Punccohar, 1994) and differences in emotional reactions to uncertain situations (Croson and Gneezy, 2009; Loewenstein *et al.*, 2001). Nevertheless, more research is needed to clearly identify the main forces driving males' behavior. Still, the difference in avatar-selection by women across treatments indicates that they were at least responsive to the perceived environment.

Regarding performance, we find strong support for the stereotype that men do better on math-related tasks: Men produced 50% more correct sums than women did. The difference is highly significant, even though it goes against some previous studies using the same task. Yet the stereotype that women are better

at recognizing emotions received no support in our data, since performance was nearly identical. In fact, we were somewhat surprised by both of these performance results. While there is typically *some* basis for stereotypes, sometimes they manifest in reality and sometimes they do not.

The patterns in the hiring decisions surprised us. Although there is an expectation in the math task that men would be more likely to be hired than women, given the widespread stereotype that males are better in this sort of tasks, there is no such overall hiring pattern and female avatars are more likely to be hired. Moreover, female avatars are more likely than male avatars to be hired in all three treatments, so discrimination against females was entirely absent in our experiment with a standard British subject pool. We do not have a compelling explanation for this phenomenon, but we do note that there appears to be a relationship between the relative scarcity of an avatar type and the likelihood that avatars of that type are hired.

Summing up, although there is a perception that men would be favored for hiring in the math task, we do not observe such an overall hiring pattern. This gap between women's expectation of how firms would behave and real behavior could reflect that people bring into the lab a definite view that women will be discriminated against in a math-related task. This belief that women will be discriminated against in *MIT* is strong enough to induce women to hide their real gender even though this was actually counter-productive. This explanation suggests a stereotype (Bordalo *et al.*, 2016) that is not related to performance, but to discrimination, that would play a role in this context and influence how subjects choose the avatar.

There has been little previous research regarding how men and women make choices when there is a perception that there could be discrimination in

hiring. Anticipation of discrimination is critical in terms of employment-related choices by workers and students intending to become workers. Potentially, our results could have interesting implications for policies aiming to reduce the gender gap in STEM careers, suggesting that it may be even more important to dispel the perception that people have about discrimination than to reduce the discrimination in the labor market, since we do not see evidence of the latter in our data.

In any case, we hope that our findings open a line towards further research on anticipated and actual gender discrimination in different environments.

2. CHAPTER 2

GROUP IDENTITY AND CHARITABLE CONTRIBUTIONS: EXPERIMENTAL EVIDENCE

2.1 INTRODUCTION

Charitable organizations exist to support and raise funds for a wide variety of causes. Apart from the organizations that are government funded, the rest depend on the generosity of the general community. Benefits from charitable contributions go from supporting these organizations that offer goods and services around the world, to also encouraging people to provide more help for those in need, which would create a society with a larger social capital and much less likely to be divisive (Charness and Holder, 2018).²⁴ Hence, encouraging philanthropic activities has become an issue of importance.

The experimental economics literature has proposed different alternatives to increase charitable giving both in the lab and in the field, such as providing information about the recipient (Branas-Garza, 2006; Eckel and Grossman, 1996); framing the donor (Branas-Garza, 2007); directly asking for a certain amount of money (Edwards and List, 2014); and providing external matching for the donations made by the participants (Karlan and List, 2007; Eckel and Grossman, 2008; Martin and Randal, 2008).²⁵

²⁴ For a discussion, see Putnam (2000).

²⁵ Note that there is also some evidence against the positive effect of external matching. For example, Rondeau and List (2008) find that providing matching funds does not increase donations in the field, and Huck and Rasul (2011) find that linear matching schemes could be counterproductive under some circumstances.

Social distance and group identity have been also explored as potential factors affecting donations. Reducing social distance both with respect to the recipient (Bohnet and Frey, 1999; Charness and Gneezy, 2008) and with respect to the experimenter (Hoffman et al, 1996) has been shown to be an effective tool to increase donations. Group identity has also been found to have an effect on subjects' behaviour. Specifically, people exhibit a greater charity and lower envy toward an in-group than toward an out-group match (Chen and Li, 2009); and group membership has strong effects on the allocations chosen for in-group and out-group members (Billig and Tajfel, 1973). Moreover, Charness, Cobo-Reyes and Jimenez (2014) show that a simple team-building exercise greatly increases the level of contribution to a Public Good regardless to whether subjects are linked to people from their team-building exercise, showing that positive feelings engendered by group identity spill over to other participants outside the group.²⁶ However, to the best of our knowledge, there is no paper that studies whether the positive spill-over effect of group identity found in PGG would also be present in a situation in which people decide on the donation to an independent third party.

This paper aims to fill that void by studying whether there is a positive effect of enhancing group identity on the amount of money donated to a charity. We take this as a starting point for our experimental design. In our experiment, subjects participate in a real effort task that consists of solving anagrams for 12 minutes. In this task there is a monetary reward for forming a sufficient number of words from a series of letters. Then, subjects move to a second activity in

26 There are other papers that study the effect of a pre-task activity on contributions to a PGG (Brown-Kruse, 1993; Cadsby and Maynes 1998; Solow and Kirkwood, 2002). However, these papers only consider the case in which participants in the PGG are matched with the same participants that they did the previous task.

which they would have to decide how much of the money generated in the first stage to donate to a charity.

We vary two factors in a 2x2 design. One factor is whether group identity is enhanced by having people to perform the anagrams task in groups or individually. The second factor concerns the procedure to donate the money to a charity (either as group or as individual donations). Note that the donation decision is always made individually, the only difference being that subjects in the “group donation” treatments are told that their individual donations are to be grouped before sent to the charity. However, in all cases the donations are decided individually and are completely anonymous.

Our results are intriguing. We find that enhancing group identity using a group task positively affects the proportion of people donating, as well as the amount donated to charity. In particular, the percentage of people not donating anything is significantly lower when the anagrams task is performed in groups (0% and 2.78% for group and individual donations, respectively) than when participants do the task individually (the respective percentages are 17.14% and 19.44%). Average donations are found to be also larger under the group task condition (£5.05 and £3.22 for group and individual donations, respectively) than when the task was done individually (the respective figures are £2.97 and £2.44).

Results also show that donating as a group has an effect on the amount donated only when group identity was previously enhanced (contributions go from 3.22 when they are made individually to 5.05 when they are grouped). However, there is no effect of the group donation if the group identity was not salient.

The closest paper to ours is that by Charness and Holder (2018) that finds that team competition for matching funds increases charitable contributions even

when the groups were randomly assigned. Although group membership plays a role in explaining subjects' behaviour, Charness and Holder (2018) focus on the effect of group identity in a team competition environment; as opposed to ours, which complements the previous findings by showing the effect that enhancing group identity *per-se* has on donations, without any sort of competition required for group identity to work. We simply look at the pure effect of group identity on charitable giving.

The remainder of the paper is organized as follows. In Section 2 we explain the experimental design. We state our main hypotheses in Section 3. Section 4 shows the main results and provides some discussion. We conclude in Section 5.

2.2 EXPERIMENTAL DESIGN AND PROCEDURES

2.2.1 Experimental design

The experimental design consists of four treatments: The *Identity-Group-Donation* treatment, the *Identity-Individual-Donation* treatment, the *No-Identity-Group-Donation* treatment, and the *No-Identity-Individual-Donation* treatment. All four treatments had two different stages. First, subjects would generate an amount of money by solving anagrams for 12 minutes (Charness et al, 2014). Then, subjects would move to the second activity in which they would have to decide how much of the money generated in the first stage to donate to a charity. We designed the experiment with the two stages to make the donation decision more meaningful. Subjects were not just endowed with a fixed amount of money that they had to divide; instead, they earned that money doing a real effort task.

In the *Identity-Group-Donation* treatment (*IGD*, hereafter), upon arrival to the lab, each of the six participants in each session picked a label from an opaque

bag containing numbers between one and six. Subjects who picked an odd number were assigned to the “orange” group, whereas even numbers were assigned to the “green” group. The room was divided into two aisles with the right side of the room being the location for the “orange” group and the left side for the “green” group.

Members of each group sat together, and each group was handed only one set of instructions.²⁷ The task consisted of creating words from a number of letters. Each group was handed an anagram and its three members had 4 minutes to solve it together. After the time was over, the experimenter collected the anagrams sheet and handed a new one. This process was repeated for a total of three anagrams.²⁸ The group would be paid £1 per correct word if the total number of words found by the group (including the three anagrams) was below 24. If 24 or more correct words were found, the group would receive a fixed amount of £54.²⁹ The money generated would be split equally among all members of the group, regardless their contribution to the anagrams task. We chose to use this payment system to make sure that there was no heterogeneity in the endowments for the donation stage. Note that during the first stage subjects did not know what the task would be they would have to perform afterwards.

After the first task was over, participants sat separately and individually received the instructions for the second part. For the second activity, subjects had to decide how much -if something- of the money generated in the first stage

²⁷ We considered that giving only one set of instructions per group would make the group task more salient.

²⁸ The reason behind having three anagrams per group instead of only one was that we wanted participants to spend at least 10 minutes together in order to enhance group identity. We considered that starting a new anagram every 4 minutes would make subjects more engaged, avoiding (or, at least, reducing) the possibility of losing interest after a few minutes.

²⁹ The level of difficulty was set to be low so every group got to receive the maximum amount of £54.

to donate to *Doctors Without Borders* (*DWB*, hereafter).³⁰ Just to be sure that all participants had, at least, a minimum knowledge about the charity, we provided some brief information about its goals and projects in the instructions (see Appendix A for the complete instructions). Subjects were also informed that the final donation sent to *DWB* would be the sum of all the individual donations in the group. To avoid peer pressure, subjects donated individually and no information regarding individual donations was released during or after the experiment.

Given the nature of the task, to ensure complete anonymity in the donation stage we followed a double-blind procedure. Each subject received an envelope and participants were told to introduce the money that they wished to donate out of their own £18.³¹ Then, they would place the envelope in a box located at the exit of the room when leaving the experiment.

The *No-Identity-Group-Donation* treatment (*NIGD*, hereafter) was similar to *IGD* with the only difference being that the subjects participated individually in the anagrams task. In this case, subjects sat randomly around the room and the three anagrams were handed individually to each participant who would be paid £0.5 per correct word if the total number of words found (including the three anagrams) was below 15 and a fixed amount of £18 for those who found 15 or more correct words. As in the previous case, we made sure that the difficulty of the anagrams was such that there was no heterogeneity in the endowments for the donation stage.

For the second stage, subjects were handed an envelope to make the donation. The process was exactly the same as in *IGD*, including the envelopes marked in orange or green. Also similar to *IGD*, subjects were informed that the

³⁰ This charity was chosen because it is generally well known to people.

³¹ To be able to identify which donation belonged to which group while maintaining the anonymity of the donor, envelopes were marked on the top corner with the color of the group (either orange or green).

final donation sent to *DWB* would be the sum of all individual donations made by members of the “group” that shared the same color and that all donations would be individual and completely anonymous.³²

After making their decision, subjects would place the envelope in a box in their way out when leaving the lab.

In the *Identity-Individual-Donation* treatment (*IID*, hereafter) the procedure was similar to *IGD* except that for the second part of the experiment subjects were explicitly informed that each individual donation would be sent to *DWB* with no reference done to a group or joint donation.

In the *No-Identity-Individual-Donation* treatment (*NIID*, hereafter) both the anagrams task and the donation were made individually.³³

2.2.2 Procedures

The experiment was conducted at the University of Exeter with 144 participants, who were recruited using the online recruitment system ORSEE (Greiner, 2015). We conducted 24 sessions (6 per treatment) with 6 subjects in each session. No subject participated in more than one session. On average, each person received £18 for a session that lasted 45 minutes or less.

To avoid people not trusting that the donations would be made to *Doctors Without Borders*, in all treatments subjects were offered the possibility of providing an email address to which they would receive an email with the confirmation that the donation was actually made through. Note that, given that the decisions were completely anonymous, the confirmation email that a subject

³² Note that in this case, as the anagrams stage was played individually, groups could be perceived as more artificial and less meaningful.

³³ However, and to keep the conditions of the donation the same across treatments, in *IID* and *NIID* treatments envelopes were marked on the top corner with one color (either orange or green) although those colors did not have any meaning for subjects in the second stage in these treatments.

would receive included the total amount transferred to *DWB* in that particular subject's session instead of his/her individual donation.³⁴

2.3 BEHAVIOURAL PREDICTIONS

In this section, we focus on subjects' behaviour in the second stage. Previous results show that group identity plays a role in increasing contributions in a Public Goods Game, even when subjects play the public goods with people different from those to which they share the identity (Charness et al, 2014). Based on these findings, we conjecture that enhancing group identity will also have a positive effect on donations to charitable organization. This will lead treatments in which the anagrams task is performed in groups to have more people donating and higher amounts donated than those in which the anagrams task was played individually, regardless on whether the donations are made individually or grouped.

Hypothesis 1.1. *A larger proportion of subjects will donate in the second stage of the experiment in IID than in NIID. This will be also the case for the comparison IGD vs. NIGD.*

Hypothesis 1.2. *Average donations in the second stage of the experiment will be higher in IID than in NIID and will be also higher in IGD than in NIGD.*

What do we predict regarding individual donations compared to group donations? The results found in Babcock et al. (2010) show that subjects are more apt to attempt an effort-intensive task when motivated as a team than when they are motivated as individuals. They conclude this is due to the evocation of

³⁴ Note that two people provided their email address in the same envelope used to donate to the charity. Given that this would violate the double-blind procedure followed in this experiment, we removed these two observations from the sample when analyzing the data.

guilt and reciprocity. Following this intuition, and despite having a different environment, we consider that a similar effect can take place when it comes to donate to charity. The evocation of guilt and reciprocity can still hold even if there is no sort of shared information about others' donation, which would lead to larger amounts donated by groups (even though donations are done individually) than by individuals.

Hypothesis 2.1. *A larger proportion of subjects will donate in the second stage of the experiment in IGD than in IID. This will be also the case for the comparison NIGD vs. NIID.*

Hypothesis 2.2. *Average donations in the second stage of the experiments will be higher in IGD than in IID and will be also larger in NIGD than in NIID.*

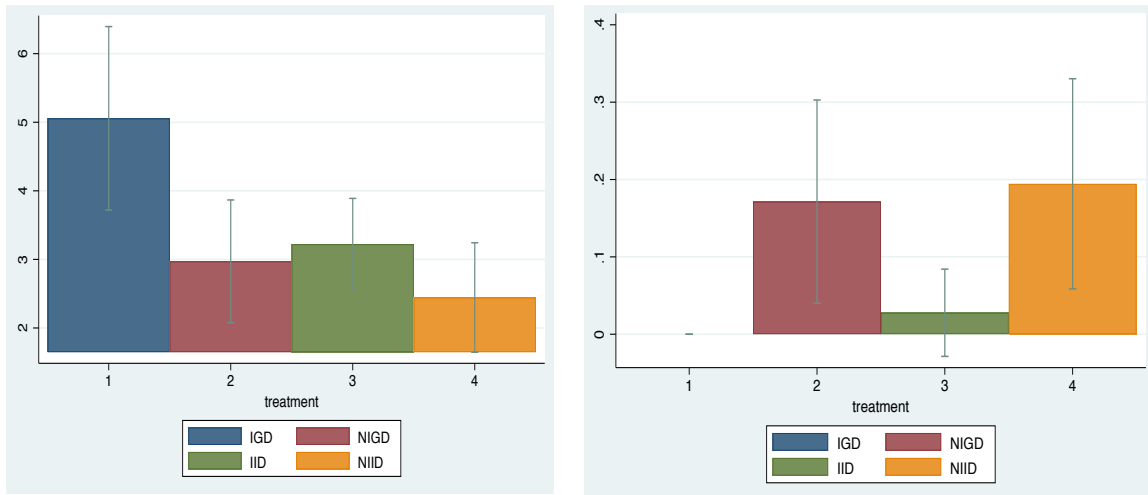
2.4 RESULTS

We first analyse the effect of group identity on two margins of charitable donations. On the extensive margin, we examine whether enhancing identity as a group works in encouraging agents to donate. On the intensive margin, we also study whether the real effort task played in groups increases the amount donated vis-à-vis the case in which the anagrams were solved individually. Second, we compare donations to a charity when they are done individually and when they are grouped. We make this comparison for two cases, when group identity was enhanced and when it wasn't salient. Note that given the nature of the experiment and the double-blind protocol followed, we do not have the donation linked to a particular individual, which will limit our analysis.

Figure 1 summarizes our findings. Figure 1(a) plots the average individual's donations (errors bars represent 95% confidence intervals) made in

each treatment. Figure 1(b) focuses on a different dimension, representing the percentage of subjects donating zero to the charity.

Figure 1. Individuals' behaviour



(a) Average individual donation

(b) Percentage of subjects donating zero

2.4.1 Effect of enhancing group identity

We begin by studying the effect that playing the anagrams in groups has on subjects' decisions regarding their donation to *DWB*. We study this effect for both the intensive (percentage of people donating) as well as the extensive (average amount donated) margin.

As shown in the right panel, the percentage of subjects who decided to donate zero to the charity when the anagrams task was done in groups is 0% and 2.78% when the donation was made as a group (*IGD*) or individually (*IID*), respectively. When the anagrams task was performed individually, the corresponding values are 17.14% (*NIGD*) and 19.44% (*NIID*). Differences are statistically significant for the case where donations were made as a group ($\chi^2 = 6.562$, $p = 0.010$, for a two-sided test for equality of proportions between *IGD* and

NIGD) as well as for the case of individual donations ($\chi^2 = 5.062$, $p = 0.024$, for a two-sided test for equality of proportions between *IID* and *NIID*). These results support our Hypothesis 1.1.

Turning into the amount donated, the left panel in Figure 1 shows that these numbers are £5.05 (*IGD*) and £3.22 (*IID*) under the group anagram task; and £2.97 (*NIGD*) and £2.44 (*NIID*) when the group identity wasn't salient. This would mean that enhancing group identity with a real effort task before the donation increases the amount sent to a charity by a significant 70% ($z = 2.640$, $p = 0.008$; Mann-Whitney two-tailed test) when individual donations were grouped; and by a significant 31.96% ($z = 1.996$, $p = 0.046$; Mann-Whitney two-tailed test) when donations were not grouped. These results are consistent with Hypothesis 1.2.

Result 1: *Enhancing group identity positively affects attitudes towards charitable donation in both the intensive and extensive margins.*

2.4.2 Effect of grouping donations

In this section, we analyse subjects' behaviour depending on whether the donation to the charity is done individually or grouped. We will study both the intensive and extensive margin of individuals' decisions.

Focusing on the percentage of people donating to *DWB*, Figure 1(b) reflects that grouping donations does not affect subjects' willingness to donate, leading to very similar percentages of people donating regardless whether there was a previous group task ($\chi^2 = 0.986$, $p = 0.321$, for a two-sided test for equality of proportions between *IGD* and *IID*) or not ($\chi^2 = 0.063$, $p = 0.802$, for a two-sided test for equality of proportions between *NIGD* and *NIID*). So, Hypothesis 2.1 is not supported by the data.

Looking at the average amount donated, results show that differences are statistically significant for the comparison between *IGD* and *IID* ($z = 2.247$, $p = 0.025$; two-tailed Mann-Whitney test) but these differences are not significant when we compare *NIGD* and *NIID* ($z = 1.014$, $p = 0.311$; two-tailed Mann-Whitney test). So, grouping donations only seems to have an effect on subjects' behaviour when there has been a previous task that makes the group identity more salient. Hence, we find only partial support for our hypothesis 2.2.

Result 2: *Grouping donations does not affect attitudes towards charitable giving on the intensive margin and only increases the amount sent to the charity when the group identity was previously enhanced.*

2.5 CONCLUSIONS

This paper analyses whether a very simple team-building exercise affects subjects' attitudes towards charitable giving. We examine the intensive and extensive margin of donation decisions using a 2x2 design in which we vary whether there was a group exercise to enhance group identity prior to the donation stage and whether donations were done individually or grouped. We find that when group identity is made salient, there is a positive effect on the attitudes towards donation from the point of view of both the intensive and extensive margin. However, grouping donations had an effect on subjects' behaviour only for the case in which the group identity was previously enhanced.

We employed a careful and clean double-blind experimental design to ensure full anonymity across subjects and with respect to the experimenter in order to avoid any sense of peer pressure. Thus, our results may well represent a lower bound for the effects we find, since peer pressure is present in most environments.

Potentially, our findings have interesting implications for philanthropic activities. Our results suggest that bringing out a (minimal) sense of community is enough to highly increase donations. Charity campaigns could benefit, without falling into higher campaign's costs, from eliciting donations from groups that share some level of identity – i.e., company, school, fraternity, etc. – as compared to promoting individual contributions.

3. CHAPTER 3

IMPROVING HEALTHY EATING IN CHILDREN: EXPERIMENTAL EVIDENCE.

3.1 INTRODUCTION

Poor diet has been identified by the World Health Organization (2009) as a major determinant of global risks to health and one important reason for the rising costs of healthcare. According to the most recent Global Burden of Disease study (2016), poor diet is linked to one in five deaths worldwide, with low intake of healthy foods being the leading risk factor for mortality. Moreover, 45% of all cardio-metabolic deaths in the U.S. were associated with suboptimal intakes of dietary elements (Micha et al., 2017).

A poor diet has been shown to not only have consequences for the adult population but to also have long-term implications for children, affecting their physical well-being and contributing to the development of dental caries, cognitive impairments, and diabetes (Noble and Kanoski, 2016). It may also weaken immune systems (Sorahindo and Feinstein, 2006), hinder growth, or undermine intellectual performance (Weinreb et al., 2002; Whitaker et al., 2006).

Recent data on children's eating patterns are far from encouraging. It has been shown that, despite the progress in fruit intake, children still fail to meet recommendations for the amount of both fruit and vegetables they should eat daily (Kim et al., 2014). A report from the National Cancer Institute (2018) states that 60% of children did not eat enough fruit to meet daily recommendations in

the period 2007-2010, and 93% of children did not eat enough vegetables.³⁵ On a different dimension, the National Health and Nutrition Examination Survey (NHANES) reported in 2011-2012 that over 30% of U.S. children between 2-19 years old were overweight. This percentage was tripled in comparison to the survey carried out in 1971-1974. Since dietary habits are typically set during childhood, it is crucial that improvements are made during this period (Haire-Joshu and Tabak, 2016).

Hence, designing, implementing and evaluating interventions aiming to produce durable changes in children's eating behaviour is critical not only to better understand the drivers and the evolution of eating habits, but also to support policy efforts aimed at tackling the long-term individual and social consequences of poor nutrition. In some sense, this paper can be viewed as a next generation of behavioural interventions, where no direct reward is paid for achieving goals. While it has been shown that paying people directly for performance leads to better performance, this is not always practical or feasible. Our approach can be implemented with relatively minimal funding and so could be widely adopted.

To attempt to address this crucial issue, we conduct an innovative field experiment in Spanish elementary schools to study whether and how it might be possible to influence or incentivize school children, aged nine to ten, to make healthier food choices. We have four treatments in our design. In all treatments, children were presented with five different food trays having an assortment of choices and made four choices from whichever trays. In the Baseline treatment, nothing else was done. In the Nutritionist treatment, a nutritionist explained the benefits of healthy eating at the beginning of the first day. In the Grades

³⁵ In Europe only 23.5% of the children eat the recommended daily amount of fruit and vegetables (Lynch et al., 2014)

treatment, a clearly-labeled “grade” was associated with each of the trays, with healthier foods receiving higher grades. Finally, in the Parents treatment, parents also received weekly information about the average food grade received by their children, and children were aware that parents would receive this information. The intervention lasted for three weeks. After four months, we returned for a surprise visit to examine long-term effects after the removal of any incentives.

Our theoretical model predicts that the Grades and Parents treatments would generate competition between participants and, based on the positive effect of competition on children's choices of fruits and vegetables found in Belot et al. (2016), we hypothesized that these two treatments would lead to more healthy choices than the Baseline. We also expected the information released by the nutritionist to significantly change the behaviour of children, so we predicted differences between the Nutritionist treatment and the Baseline.

In fact, our results show that there is a modest improvement in healthy eating in both the Grades and Nutritionist treatments, although there are different trends over time. For example, the proportion of healthy choices is 36% in the Baseline, 46% in the Grades treatment, and 45% in the Nutritionist treatment. The increase in the Parents treatment is quite dramatic, with 74% of the choices being healthy foods; note that this rate is more than double the rate in the Baseline. We show that this is not due to outliers. Perhaps most critically, we also show that these effects persist over time even after the removal of incentives. In “surprise” sessions conducted four months later (and without any stimulus), the respective proportions of healthy choices are 41%, 54%, 47%, and 69%, depending on the form of the incentives that had been in force four months before. We observe on average a positive change in the rate of healthy choices

made compared to the rate when the intervention was in force. It certainly seems that there are benefits to these interventions even after their removal.

The contribution of this paper is threefold. First, this paper proposes a sustainable and almost costless way of improving healthy choices in children. We show that engaging parents as change agents and providing them (with the knowledge of their children) with information about children's food choices is quite effective. In the previous literature, the role of parents has been primarily restricted to two strategies to influence children's food intake: restriction and pressure. Evidence has shown that pressure to eat a target food often influences the preference for the food negatively, and that restriction may increase the desire and subsequent intake of the restricted item (DeCosta et al., 2017). In our study, we incorporate parents into the decision process in a non-invasive way. Parents receive information about the average grade obtained by their children, but do not have exact information regarding which specific items they chose. So, while parents can still try to influence or stimulate their children's decisions, the final choice is on children, who have the flexibility to choose the food they prefer without any specific restrictions, since they make their choices in school.

Second, as mentioned above, the observed improvements in diet are on average fully present four months after the intervention is over and the incentives have been removed. Our results are even more striking since the surprise visits were conducted not only four months after the first intervention but also within a different academic year. In a sense, our approach of incentivizing children seems to generate some sort of good habits or learning, shedding some light on how to achieve long-term impacts. As far as we know, only material incentives have been shown to be able to generate a post-intervention effect. Yet even this evidence is inconclusive: While List and Samek (2015) and Loewenstein et al. (2016),

discussed in more detail in the next section, do find that the percentage of healthy choices is larger than in the baseline conditions, it drops significantly after the end of the intervention. In Belot et al. (2016), there is no significant long-term effect once the material incentives have been removed.³⁶

Finally, this paper also contributes from a methodological perspective. We provide useful (grading) information to children who are familiar with the grading system used to assess their schoolwork, and who will thus be able to clearly understand what it means for their food choices to be satisfactory or not – i.e., to “fail” or “pass” – given the marks they receive. This feature relates our paper to those studies that focus on choice architecture and nudging. Most of these studies generally use labels that either simply try to make the food attractive to subjects (Morizet et al., 2012; Pelchat and Pliner, 1995) or provide nutritional information that requires a high level of literacy and numeracy to interpret (Rothman et al., 2006). These interventions have been shown to have either little or no effect at all on choices.

An alternative approach has been to use colour codes that classify food items into red, yellow and green depending on their healthiness. This approach shows only slightly better results than the baseline in a hospital cafeteria (Thorndike et al., 2012), using weekday data. There is no evidence that any colour code interventions have had longer-term effects. Our information system has several advantages over the colour code. The grade system provides more flexible and accurate information about the nutritional value of the food while retaining simplicity. This information makes it easier for people to balance consumption when they choose more than just one item. Finally, the proposed

³⁶ In a different dimension, Charness and Gneezy (2009) in their exercise study find that good habits are sustainable for at least some period after a successful intervention.

system makes the general information about choice and consumption easier to understand and use for other potential agents involved in the process.

We feel that our intervention, particularly the treatment involving parents, offers great promise for successfully battling the problems of poor childhood diet and obesity that has been expanding in the developed world. Changing the eating habits for school children is a critical problem to tackle: If sustainable healthy eating habits are established early, we should expect less obesity in the adult population as well. In addition, it may well be the case that parents become inspired by the successful dietary change made by their children and also adopt a healthier diet. All in all, we feel that this is a very promising avenue.

The structure of the remainder of this paper is as follows. Section 2 offers a literature review of previous related work, while section 3 describes our experimental design in detail. We present our hypotheses in section 4 and results in section 5. Finally, we discuss these results and conclude in section 6.

3.2 LITERATURE REVIEW

As was discussed briefly in the introduction, this paper relates mainly to two streams of the literature. First, our experiment links to studies that analyze the effect of parental control on food choices of children, although our approach is radically different from what has been done previously. Second, this paper also connects to those that have investigated the effect of providing food information on healthy decisions.

Regarding parental control, we mentioned two main strategies that have been used by parents to influence children's behavior: restriction and pressure. Restriction has been found to have negative consequences in terms of eating behavior; prohibition leads to increased desire and consumption when the

forbidden food becomes available (Fisher and Birch, 1999; Jansen *et al.*, 2007; Jansen *et al.*, 2008). Ogden *et al.* (2013) also found that the restricted group was more preoccupied with the target food. Thus, we are not aware of any studies indicating that restriction is an effective strategy for effecting dietary change.

Galloway *et al.* (2006) tested the effect of pressure to eat on food intake in children. In their study, children had to eat soup under two conditions: In the pressure condition, children were reminded to “Finish your soup, please”, four times during the session. In the no-pressure condition, children were not pressured to finish their soup. They find that the increase in intake was *higher* in the no-pressure condition. Rigal *et al.* (2016) studied the effect of harsh vs. gentle instructions on food acceptance. Children were exposed to baby corn under two conditions: gentle (“You may eat that food. Try to taste it. It’s good”) or harsh (“You have to eat that food”). Results show that change of intake was higher in the gentle condition.

Our study, although related to the previous literature, offers a completely different approach of parents’ involvement in children’s decision process. Parents neither can restrict the food options (since choices are made by children in school) nor can they put pressure on children when they are making their decision. In our setting, although parents can still try to influence or stimulate their children’s decisions (given that they receive weekly information about the average grade obtained by their children), the final choice (made in school) is by the children, who have the flexibility to choose the food they prefer without any specific restrictions.

Attention has been paid to the role of the provision of information in improving healthy choices. Wisdom *et al.* (2010) studied the effect of calorie

information and “asymmetric paternalism” on the type of food picked in a chain restaurant. They find that providing the calorie information had no significant effect on the probability of picking a low-calorie sandwich. However, participants were more likely to pick it in a paternalism treatment in which it was more convenient to pick the low-calorie sandwich. Along the same line, Roberto *et al.* (2010) studied the effect of calorie labels on food choices in a restaurant. While there was a short-term effect of the calorie labels on consumption, results show that calorie labels had no effect almost immediately after they were no longer in place.

There is no robust positive effect of calorie labels on healthy choices (see Downs *et al.*, 2013, and Roberto *et al.*, 2010), which could reflect the fact that nutritional information is not always easy to understand, and many nutrition labels require a high level of literacy and numeracy to interpret (Rothman *et al.*, 2006). To tackle this problem, scholars have proposed an alternative way of providing information about the food that would be easier to understand, showing more promising results. For example, Vyth *et al.* (2011) investigated (in Dutch cafeterias) whether labeling foods with the Choices nutrition logo (a logo introduced in several large catering organizations in the Netherlands in 2006) had an effect on food choices. While they find a significant effect of the logo on fruit sales, no significant differences were found in sales for the other items (soup, bread, salad and snacks).

In the same vein, Thorndike *et al.* (2012) introduced a color-code system in a Massachusetts General Hospital cafeteria over 3 months, using the labels red, yellow and green to label unhealthy, less unhealthy and healthy items, respectively. The sales of red items decreased by 9.2% and green ones increased by 4.5%, with the results coming mainly from the effect of color labels

on beverage choices. Results are less conclusive, though, when authors perform a difference-in-differences analysis between the cafeteria and two comparison sites. Results from between-groups differences are also inconclusive. Choices in the intervention site were healthier than those in the comparison sites for some items, while for other differences were small and insignificant (and sometimes even less healthy). Thorndike *et al.* (2014) study the long-term effect of the color labels system proposed in the earlier study. Two years of keeping the labels and positioning of foods to make healthy items more accessible and unhealthy items harder to get led to a modest shift to green items. Critically, however, the incentive is kept in place for the *entire duration* of the study (unlike our design), meaning that it cannot speak to the potential persistence of effects after the end of the intervention.

Finally, although not directly related to our paper, the first generation of interventions in behavioral economics provided material incentives aiming to promote healthy behavior.³⁷ These incentives have been shown to have an effect in the short-term, but the long-term effect is not yet conclusive. For example, Just and Price (2013) conducted a field experiment at fifteen elementary schools in Utah. They assign students from the different schools to different incentive schemes in which they change the reward for eating a serve of fruit or vegetables per day (a quarter or a prize) and the time the child receives the reward (immediately or in one month). There was a 27% increase in the fraction of children eating at least one serving of fruit or vegetables when

³⁷ Material incentives have been used to promote healthy behaviors other than the one addressed in this paper. For example, Charness and Gneezy (2009) demonstrate that financial incentives can have long-lasting positive effects on individuals' willingness to exercise that persist even after such incentives are removed. John *et al.* (2011) and Volpp *et al.* (2008) provide evidence of the effectiveness of financial incentives for weight loss. Volpp *et al.* (2009) show that financial incentives significantly decrease the smoking rate, although only during the intervention. Of course, one problem with monetary incentives is that they are difficult to roll out on a large scale. Another is that this can become quite expensive and most likely monetary incentives are impractical in the long run with the general population.

any incentive was offered, but the authors do not study the effect of the short-term rewards on the long-term behavior, which could potentially even go in the opposite direction through crowding-out of intrinsic motivation (Deci and Ryan, 1985).

Belot *et al.* (2016) conducted a field experiment in schools in England to test the effectiveness of two alternative incentive schemes on choosing fruit and vegetables: a piece rate and a tournament. In both schemes, participants got a sticker for choosing a fruit or vegetable. In the piece-rate treatment, subjects who collected four stickers over the week would get a prize. In the tournament treatment, only the child with the largest number of stickers would get the prize. Results show that there was little effect of the piece-rate scheme on choices but there was a positive effect of the competition mechanism. However, choices were not significantly different in the long term once the incentives were removed.

List and Samek (2015) performed a field experiment in Chicago in which children were given a choice between a dried fruit cup (healthy item) and a cookie (unhealthy item), and they were allowed to select only one item. There were four treatments: i) a gain-frame incentive (the child received a small prize for consuming a fruit cup), ii) a loss-frame incentive (the child received a small prize that was taken away if one did not consume the fruit cup), iii) a 3-min educational message about the benefit of fruits versus cookies, and iv) a loss-frame incentive combined with the educational message. Paying cash directly for performance was effective: The proportion of children choosing the healthy snack increased from 17% in the baseline to nearly 80% when the incentives were introduced. But do the effects continue when the flow of money is cut off? While there is some evidence of post-intervention effects, the data are from only

one week after the incentives were removed and can thus only be considered suggestive evidence.

Loewenstein *et al.* (2016) conducted a field experiment in elementary schools in Utah. During the reward period, children eating at least one serving of fruits or vegetables received a token worth 25 cents that could be redeemed at a school store, school carnival, or book fair. The incentives were in place for either three or five weeks. They find a strong positive effect of the reward on the percentage of children eating a serving of fruits or vegetables: the percentage of children that ate at least one serving of fruits or vegetables increased from 38-40% in the two treatments to 76-80%. However, two months after the incentives were removed, this percentage drops dramatically (to 48-54%), although it is still larger than in the baseline condition. Thus, roughly 70% of the increase from the baseline had already vanished during the two-month period after the intervention.

Finally, Belot *et al.* (2019) conducted a field experiment to examine whether dietary habits are malleable. Low-income families in the UK participated in one of two treatments. In the “Meal” treatment, families received free groceries and were asked to cook five healthy meals per week. In the “Snack” treatment, families were asked to reduce snacking and eat at regular times. The two treatments were implemented for 12 weeks. Results provided evidence that children in both treatments reduced their body mass index compared to children in the control group. They also found some evidence that sugar intake was reduced in both treatments. However, there was no strong evidence that children’s preferences changed in favor of healthier foods. One potential explanation for the patterns observed may be that the interventions

had an impact on what the parents fed their children, rather than on children's preferences. In our study, we observe and target the latter.

3.3 EXPERIMENTAL DESIGN AND PROCEDURES

3.3.1 Experimental design

Our experimental design consists of four different treatments:

- *Baseline*: Children participating in the experiment were presented with five different food trays. Each tray included five different food items of similar nutritional value, selected by a nutritionist.³⁸ Subjects had to pick four food items from any combination of the trays. They could choose items from the same or from different trays and were able to select more than one item from the same tray and more than one unit of the same item.

- *Nutritionist Treatment (NT, hereafter)*: Similar to *Baseline*, but a nutritionist gave a short talk explaining the benefits of healthy eating on the first day of the experiment (before children made their decisions). Note that the nutritionist gave a general talk for everyone and did not lead children to pick any particular item from the trays.³⁹

- *Grades Treatment (GT, hereafter)*: Similar to *Baseline*, except that students saw labels with a "grade" associated with each of the five trays before making choices. The grades corresponding to each tray depended on their nutritional content and were analogous to those used to mark children's academic schoolwork (in Spanish schools the grades range from 0 to 10, so the five trays had assigned marks of 0, 2.5, 5, 7.5, and 10). All the items in a given tray had the same grade assigned.⁴⁰ Children were just shown the grades associated with

³⁸ See Appendix A for the nutritionist's justification of the allocation of food items to trays.

³⁹ A summary of the nutritionist's talk is reported in Appendix B.

⁴⁰ See Appendix C for the detailed composition of the trays and the corresponding grade associated with each one.

each tray and were told that these grades represented the nutritional value of the tray. No extra information was given to participants in this treatment.

- *Parents Treatment (PT, hereafter)*: Similar to *GT*, with the difference that parents received information about the average mark (linked to the nutritional composition) of their child. Before selecting the items, children were aware of the fact that their parents would receive a report about their “performance”. Note that parents did not receive exact information about the choices of their children, but rather just the average grade received over the past week.

3.3.2 Procedures

The experiment was conducted in 12 Spanish schools that were participating in a European Union (EU) program aiming to encourage healthy eating at schools. Our experiment was run as an additional activity within the EU program in these schools, so that children would not perceive their decisions as artificial, mitigating potential concerns about the external validity of our findings. Schools participating in the EU program received, throughout the academic year and depending on the week, different types of fruits and vegetables to be distributed among children.⁴¹ Students received the corresponding food, and their only choice was whether to eat it. We felt that the children in the program would see the experiment as an alternative activity in which the main difference is that they could choose the food they preferred to consume.

The selected schools were chosen to produce a large and diverse pool of potential subjects, as measured by the demographic and socio-economic characteristics of the students and their family background.⁴² Each participating

⁴¹ In 2018, 79,000 schools across Europe involving over 30 million children participated in this program.

⁴² Note that every school that was approached agreed to participate in the study. This should help to mitigate potential concerns about some sort of self-selection bias.

school was randomly assigned to one of the treatments described above. This means that all participants from the same school faced the same incentive during the intervention.⁴³ One class was randomly selected from each school to participate in the experiment. The total sample of 282 students between the ages 9 and 10 was almost evenly distributed over the 12 classes.

On the first day of the intervention in each school, experimental subjects belonging to the same class were gathered in a room (Room A). Participants were not given any information about the experiment. Each student was then – independently and sequentially – asked to move to another room (Room B) in the company of one of the experimenters. In this second room, the subject was walked through a short orientation session – the details of which depended on the subject’s treatment assignment. Next, the student picked the four food items he or she preferred. Finally, the student was taken to a third room (Room C), where he or she joined other classmates who had already completed the task. This procedure ensures that the decisions of subsequent classmates (who remained either in Room A or C at the time the participant was choosing her lunch in Room B) were not directly influenced by the decisions made by previous individuals.

A member of the research team, who was present during the decision process, recorded the students’ dietary choices and grades. Note that in order to minimize interaction with the other participants, children made their decisions alone. The bags used to keep the food were transparent so the researcher could see what children picked, without being directly involved during the decision. The

⁴³ We are not too concerned about contamination across schools, since children typically went directly home from school and in any event were quite young and thus highly unlikely to socialize with children from other schools.

intervention was always conducted right before the main school break.⁴⁴ To make the food decision salient, in every treatment the parents were instructed not to prepare any snack for their children for the days of the intervention.⁴⁵ In this way, children chose the food they would eat during the break that day. A staff member of each school was present in Room C, where children would gather together after making their choices. This person made sure that no food was wasted (although children could keep the food for later consumption). As a result, most of the items chosen in Room B were consumed in Room C.⁴⁶

To analyze the dynamics and long-term effects of the incentives, we collected data twice a week for three weeks and so have six observations per subject.⁴⁷ Importantly, to keep the same conditions across treatments, the consent letter was the same in each treatment and the parents received exactly the same information about the experiment and were asked to provide their contact details before the experiment began. Note that parents in *PT* only received information about their child's average grade in a given week. They did not receive any other information about other children's performance. The information was released once a week and included the average grade of the particular child for each of the two sessions ran during that week.

After the sixth day of the experiment children were given a questionnaire that asked them to identify their closest friends in class. It is conceivable that participants would talk to each other after the experimental session and discuss their choices. Hence, information about classmates' decisions might affect a

⁴⁴ This break is at 11am. Typically kids go outside during this break, play some games and have a snack to help them keep going until the lunch break that – in Spain – only starts at 2:00-3:00pm.

⁴⁵ While we could not enforce that children didn't bring any food to school, we ensured that they did not have access to their backpacks from the moment that they participated in the experiment until after the school break was over.

⁴⁶ Note that we did not find any differences across treatments in the amount of food saved for later consumption.

⁴⁷ The procedures to collect the data in subsequent days were the same as in the first day, i.e., participants would be located in different rooms and decisions would be made privately.

child's choices over the course of the intervention and so we tried to elicit each subject's social network in order to account for these potential interdependencies.

We also gave post-experimental questionnaires to children, teachers and parents, in order to obtain information about participants' socioeconomic variables, self-control indicators, and eating habits. Regarding the socioeconomic background of the student, we utilized a questionnaire administered by the Spanish Government and filled by the parents. The set of variables includes i) the highest educational level achieved by the parents, ii) proxy variables for the economic level of the household, and iii) parents' involvement in the school-related activities (homework) of their child.⁴⁸ For the self-control indicators, following Tsukayama *et al.* (2013), children answered questions related to their level of interpersonal and schoolwork impulsivity.⁴⁹ For eating habits, parents at home filled out a short survey designed by a nutritionist. We collected information regarding the weekly consumption of a variety of food items belonging to five different food groups (milk and dairy products, carbohydrates, proteins, fruit and vegetables, and fats and sugars).⁵⁰ In the teacher questionnaire, we gathered information regarding students' average performance in class, average attendance to the school, and a measure of self-control.

Table E.1 in Appendix E presents descriptive statistics for variables capturing children's dietary habits and background characteristics, built from the responses to the post-experimental questionnaires. Covariate balance checks reported in Tables E.2 and E.3 in the Appendix reveal no systematic imbalances in either the dietary habits or the socio-economic characteristics of the subjects assigned to the four treatment arms, indicating that randomization was

⁴⁸ See Appendix D.1 for the complete questionnaire.

⁴⁹ See Appendix D.2 for the complete questionnaire.

⁵⁰ See Appendix D.3 for the complete questionnaire.

successful.⁵¹ Nonetheless, some of the regression models reported in Section 5.4 incorporate covariates in order to account for any remaining imbalance for which we did not control in our experimental implementation (Raab and Butcher, 2005).

Finally, four months after the end of the intervention and during the next academic year, all subjects participated in a *surprise* session, which was conducted as a one-shot *Baseline* for everyone and which was unannounced.⁵² The aim of this was to study whether the effect of the different incentives remained after some time and once incentives had been removed.

3.4 Theoretical predictions and hypotheses

In this section, we offer a simple model that captures the main features of our experimental design and that allow us to state our main hypotheses.

Consider i 's utility function to be:

$$u_i(x_i) = -(x_{it} - x_{pref})^2 - \delta_{it}\alpha_i(x_{it} - x_{norm})^2 + \gamma_{it}\beta_i(g_{it} - \bar{g}_{j,t-1})$$

where x_{it} represents the amount of healthy food consumed by subject i in period t and $x_{it} - x_{pref}$ captures the difference between i 's healthy food consumption in period t and i 's preferred healthy food consumption.

⁵¹ The covariate balance checks reported in Table E.2 account for the clustered nature of our randomization using the methodology developed by Hansen and Bowers (2008), based on *Fisher's* randomization inference. For robustness, in Table E.3 we fit a multinomial logit model for treatment assignment (e.g. Gerber *et al.* 2009), using Ibragimov and Muller's (2010) approach to modelling clustered data while accounting for the small number of schools in our sample (see also Esarey and Menger, 2019). The conclusions emerging from both tables are similar, highlighting the lack of systematic covariate imbalances across treatment arms.

⁵² Note that parents were also unaware of this surprise session meaning that we could expect that, unlike the previous sessions, parents would pack snacks for this day. However, we consider that this should not affect children's decisions since, similar to the rest of the experiment, students did not have access to their backpacks until the end of the school break which, combined with the fact that children cannot eat anything during class, would mean that they could not eat the snack brought from home until school ended.

Basically, what the utility function proposes is a satiation point. Subject i wants to consume exactly his or her preferred amount of healthy food. Deviating from that preferred level of healthy consumption reduces i 's utility. Consider, for example, $x_{pref} = 0$, so that i 's preferences are such that subject i wants to consume only junk food. So, the healthier food subject i consumes, the lower her (short run) utility will be.

In what follows below, we define the following:

- $x_{it} - x_{norm}$ captures the difference between i 's consumption in period t and i 's expectation of the social norm regarding healthy consumption. Note that here we assume that subjects experience some disutility – represented by α_i – from deviating from what they consider to be socially “good”,
- δ_{it} represents the erosion factor of α_i over time. This factor considers that subject i will care more about the social norm at the beginning but over time he or she starts potentially caring less about behaving as the society dictates,⁵³
- $g_{it} - \bar{g}_{j,t-1}$ shows the difference between i 's average grades and the observed average grade of other people in i 's group in the previous period. This difference captures a form of competitive preferences – represented by β_i – in which subject i will increase her utility if she is getting a higher grade than her colleagues in class. As explained in the experimental design, grades are a function of consumption and will play a role only in *GT* and *PT*,
- γ_{it} represents the erosion factor of β_i over time.

To compute the optimal consumption in *Baseline*, given that grades are not announced and that the optimal consumption of healthy items is

⁵³ It is also possible that over time a child learns better about the social norm, so the norm itself would change.

not explicitly defined, we assume that $\beta_i = 0$. Hence, the optimal consumption of healthy items in *Baseline* would be:

$$x_{it}^* = \frac{2x_{pref} + 2\delta_{it}\alpha_i x_{norm}}{2 + 2\delta_{it}\alpha_i}$$

NT differs from the *Baseline* in that the socially accepted amount of healthy food is more accurately defined. In this case, it will be the nutritionist who states what is the desirable consumption of healthy items. While previous literature is not conclusive regarding the effect of providing information about diet, we did indeed suspect there would be some effect in *NT*, given that the information came from an authority figure and an acknowledged expert. While we thought the effect might not be quite as strong or as long-lasting (given that the nutritionist visited only one time) as in the other non-*Baseline* treatments, we have:

Hypothesis 1: *There will be a higher proportion of healthy items picked in NT than in the Baseline.*

In *GT*, as noted in the experimental design, food is graded according to its healthiness. Given that grades are directly correlated with food choices, we replace g_{it} with x_{it} and $\bar{g}_{j,t-1}$ with $\bar{x}_{j,t-1}$ – the average proportion of healthy choices by people other than i in period $t-1$. In this case we allow for $\beta_i \geq 0$. The optimal consumption in *GT* would be

$$x_{it} = \frac{2x_{pref} + 2\delta_{it}\alpha_i x_{norm} + \gamma_{it}\beta_i}{2 + 2\delta_{it}\alpha_i}$$

Following Belot *et al.* (2016), who find a positive effect of a competition mechanism on children's choices of fruits and vegetables, we assume that $\beta_i > 0$ even though our environment does not incentivize competition directly. A

strictly positive β will lead to a higher proportion of healthy choices in *GT* compared to *Baseline*.

Hypothesis 2: *There will be a larger proportion of healthy food chosen in GT than in Baseline.*

Regarding subjects' behavior in *PT*, we conjecture that the presence of parents will make grades even more important since it allows for the possibility of also getting social recognition at home. Experimental literature has found status seeking to be strongly related to competition among individuals (Huberman *et al.*, 2004; Rustichini and Vostroknutov, 2008; Charness *et al.*, 2014). Based on these findings, we assume that having the grades released to the parents would make β_i larger in *PT* than in *GT*, leading to a higher proportion of healthy items chosen in *PT*.

Hypothesis 3: *There will be a larger proportion of healthy food chosen in PT than in GT.*

Regarding the dynamics of the experiment, we assume that the erosion factor in our model will play a role, reducing the effect of incentives over time while the incentives are still in place, i.e., over the six experimental days across the three weeks of the intervention. We base our assumption on the lack of a long-term effect found in Belot *et al.* (2016). This will lead to our fourth hypothesis.

Hypothesis 4: *The proportion of healthy food chosen in all treatments will decrease over time during the intervention period, while incentives are still in place.*

3.5 Results

This section is structured as follows. We begin by comparing the average behavior of children in each treatment. We then examine the dynamics and evolution of decisions over time during the intervention period and persistence effects in the longer run. Finally, we explain the main determinants of subjects' choices.

3.5.1 Children's behavior

We start with an overview of the decisions made by the participants in each treatment. Table 1 presents a summary of the average individual grades, and the proportion of healthy choices made by individuals. For the second analysis, we define as healthy choices those items that were assigned a grade of 7.5 or above.⁵⁴

As seen in Table 1, the average grades in *NT* (5.50) and in *GT* (5.63) are both larger than in *Baseline* (4.77). Furthermore, the grades in every session of *NT* and *GT* are higher than in any session of the *Baseline*, so we have the highest possible significance level when comparing *Baseline* to either *NT* or *GT* ($p = 0.050$ for the one-tailed Mann-Whitney tests of our directional hypotheses, using

⁵⁴ As a robustness check, we replicated this analysis for the case in which we define as healthy choices those items that were assigned a grade of 10 and for the case in which healthy choices are considered those with a grade of 5 and above. Results remain qualitatively similar in both cases (see Table E.4 in Appendix E).

session-level data as independent observations).^{55,56} There is no significant difference between the averages in *NT* and *GT*.⁵⁷

Table 1. Average grades and proportion of healthy food choices for each treatment

Treatment	Observations	Subjects	Average grade	Proportion of healthy choices
<i>Baseline</i>	400	73	4.77 (2.09)	0.36 (0.26)
<i>NT</i>	412	71	5.50 (1.99)	0.45 (0.27)
<i>GT</i>	373	65	5.63 (1.85)	0.47 (0.25)
<i>PT</i>	432	73	7.88 (1.61)	0.74 (0.23)

Notes: Standard deviations are reported in parentheses.

So, it seems that simply providing information to children, either by having grades assigned to the food or by having a nutritionist explaining children the

⁵⁵ If we were to assume that each child's decision was independent the *p*-values would be as follows: *Baseline* vs. *NT*: *p* = 0.005; and *Baseline* vs. *GT*: *p* = 0.000. Yet, we prefer to report more conservative *p*-values that take into account potential dependencies of data within a given class. Note that all *p*-values reported here are rounded to the nearest third decimal place.

⁵⁶ Alternatively, some have argued that *t*-tests should be preferred to Mann-Whitney tests in very small samples, as they perform well even with as few as two observations per group (Janusonis, 2009; Winter, 2013). Other authors (e.g., Ludbrook and Dudley, 1998) suggest using permutation (exact) tests instead when sample sizes are small. The main conclusions drawn from the Wilcoxon ranksum tests remain unchanged using either *t*-tests (*t* = 7.335, *p* = 0.001, one-sided, for the comparison between *GT* and *Baseline*; *t* = 7.652, *p* = 0.002, for *NT* versus *Baseline*) or permutation tests (*p* = 0.047, one-sided, for *GT* versus the *Baseline*; *p* = 0.098, two-sided, for *NT* versus *Baseline*).

⁵⁷ Differences are not significant when we compare average grades in *GT* and *NT* (*Z* = 0.218, *p* = 0.414, one-tailed Mann-Whitney test). The conclusions are again the same using *t*-tests (*t* = 0.241, *p*-value = 0.411, one-sided) or permutation tests (*p* = 0.393, one-sided).

benefits of eating in a salubrious manner, does help to improve healthy food choices in this part of our study. Moreover, perhaps surprisingly, the two ways of providing information seem to have a very similar effect on subjects' decisions. These results support both *Hypothesis 1* and *Hypothesis 2*.

Differences in grades are dramatically larger when information is released to the parents. Results from Table 1 show that the average grade in *PT* (7.88) is far higher than in any other treatment. Once again, the (one-tailed) Mann-Whitney test on school-level data gives the highest possible significance level for each pairwise comparison to the *PT* treatment ($p = 0.05$ in all cases).⁵⁸ These results support *Hypothesis 3* and thus the idea that involving parents in the process by making them aware of their children's decisions is a very strong mechanism for encouraging healthy behavior.

Turning to the proportion of healthy choices, results are very similar to those using the average grades. As shown in Table 1, the percentage of healthy items chosen in the *Baseline* is 36%. This percentage increases in *NT* (45%) and *GT* (46%) and more than doubles with *PT* (74%) compared to the *Baseline* rate.

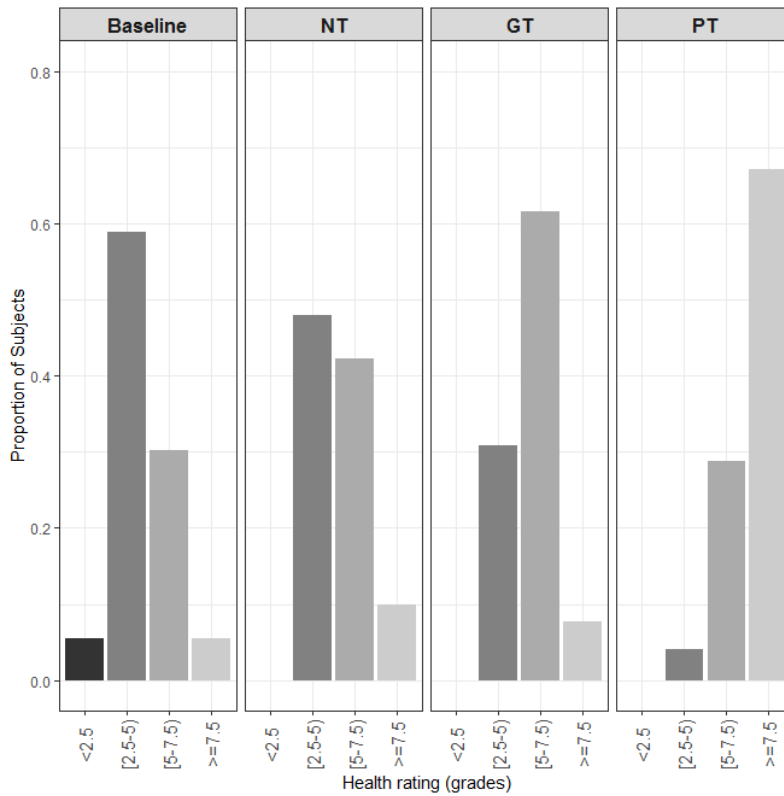
These findings support the previous results, showing that there is a moderate effect on choices in *NT* and *GT*, and that children radically change their behavior in *PT*, where the parents are informed.

⁵⁸ Similarly, $t = 16.07$, $p = 0.000$ for *PT* versus *Baseline*; $t = 7.335$, $p = 0.001$ for *PT* versus *GT*; and $t = 7.652$, $p = 0.001$ for *PT* versus *NT*. The p -values of the corresponding permutation (exact) tests are all below 0.05. School-adjusted t -tests using individual-level data and accounting for cluster randomization (Donner and Klar, 2000) also give very strong results for the comparisons between *PT* and each of the other three treatments: $t = 15.879$, $p = 0.000$ versus *Baseline*, $t = 7.217$, $p = 0.001$ versus *GT*, and $t = 7.615$, $p = 0.000$ versus *NT*.

Result 1: *Providing information to children regarding the nutritional value of the food (GT and NT) has a moderate, but nevertheless significant effect on choices. Releasing the information about the child's average grade to her parents (in PT) has a very strong effect on subjects' decisions, largely increasing the consumption of healthy food items.*

We now investigate a potential explanation for the reported positive effect that releasing the grade information to the parents has on children's choices. We argue that the average increase in healthy choices is generated by the effect that the incentive has on the majority of the population, however we could also observe the same results if there were a stronger effect affecting only a fraction of the subjects. To study the extent to which the majority of subjects is affected by the incentive, we compute the average grade of each individual and then allocate this to one of the following categories, with the average grade: *i)* at or below 2.5, *ii)* between 2.51 and 5, *iii)* between 5.01 and 7.49, and *iv)* at or above 7.5. Figure 1 plots the proportion of subjects who fall into each of the four choice categories.

Figure 1. Proportion of subjects in each category, by treatment



We see that the distribution of subjects substantially changes in *PT* compared to all the other incentive schemes. Almost 70% of the population in *PT* had an average grade of 7.5 or higher. This percentage is significantly larger than in *Baseline* (5.5%), *GT* (12.20%) and *NT* (9.85%).⁵⁹ Moreover, the distributions in *GT* and *NT* are very similar (two-sample Kolmogorov-Smirnov test: $D = 0.21$, $p = 0.190$, two-tailed). These results suggest that the improvement in healthy choices found in *PT* comes mainly from the effect that the incentive had on the majority of participants rather than from a very strong effect on just a subset of the population.

⁵⁹ All differences are statistically significant at the 5% level with (one-tailed test) Mann-Whitney tests on school-level data using *PT* as the reference group. These differences with respect to *PT* are significant at the 1% using a (one-tailed) t-test. Cluster (school)-adjusted chi-square tests for individual-level data yield similar conclusions: $\chi^2 = 4.864$, $p = 0.027$ versus *Baseline*; $\chi^2 = 4.909$, $p = 0.027$ versus *GT*; and $\chi^2 = 4.939$, $p = 0.026$ versus *NT*. Note that, since the outcome here is binary, cluster-adjusted chi-square tests are appropriate. If we ignore the dichotomous nature of the outcome and we use cluster-adjusted, one-tailed t-tests, we have: $t = 12.548$, $p = 0.000$ versus *Baseline*; $t = 9.598$, $p = 0.000$ versus *GT*; and $t = 11.427$, $p = 0.000$ versus *NT*.

Result 2: *A majority of participants react to the incentives in PT, leading to the large improvement on healthy choices.*

3.5.2 Dynamics during the intervention period

This section focuses on the dynamics in the four treatments. Figure 2 represents the average percentage of healthy choices over time, i.e., over the six experimental days across the three weeks of the intervention. Note that, similar to Table 1, Figure 2 considers as healthy choices items that were assigned a grade of 7.5 or 10.⁶⁰ We observe two clear patterns in our data. While Figure 2 shows a positive trend in *GT* and *PT*, this trend is clearly negative in *Baseline* and *NT*.⁶¹ So, the percentage of healthy choices increases over time in *GT* and *PT* but drops over time in *Baseline* and *NT*.

This result only partially supports *Hypothesis 4*. One intuition for this result is that the grades introduce competition among the children. Note that even if the decisions were made individually in an isolated room, children were gathered together after choosing the food. They thus had information about others' choices, allowing them to compute the grades of others. This could influence subjects' decisions in subsequent periods, which would lead to an improvement in choices over time. Competition has been shown to have strong effects on behavior.⁶²

⁶⁰ Figure E.1 in the Appendix provides the same analysis considering the average grade and the percentage of people with an average grade >7.5 as variables.

⁶¹ All trends are statistically significant ($Z = 2.68, p = 0.007$; $Z = -2.72, p = 0.006$; $Z = 2.68, p = 0.007$; and $Z = 4.99, p < 0.001$; Cochran-Armitage test for *Baseline*, *GT*, *NT*, and *PT*, respectively).

⁶² Teachers informed us that children in class were comparing their grades and talking about how to improve the next day, so we do think that competition played a role here, as it appears to have in Belot *et al.* (2016). Examples of the effect of competition include Bornstein, Gneezy, and Nagel (2002), who find that competition between groups increased contributions in the minimum-effort game, Fershtman, Gneezy, and List (2012), who find that just having the notion of competition in the air led to social preferences vanishing, and Charness and Holder (2019), who show that competition between groups to have their charitable

Figure 2. Percentage of healthy choices over time

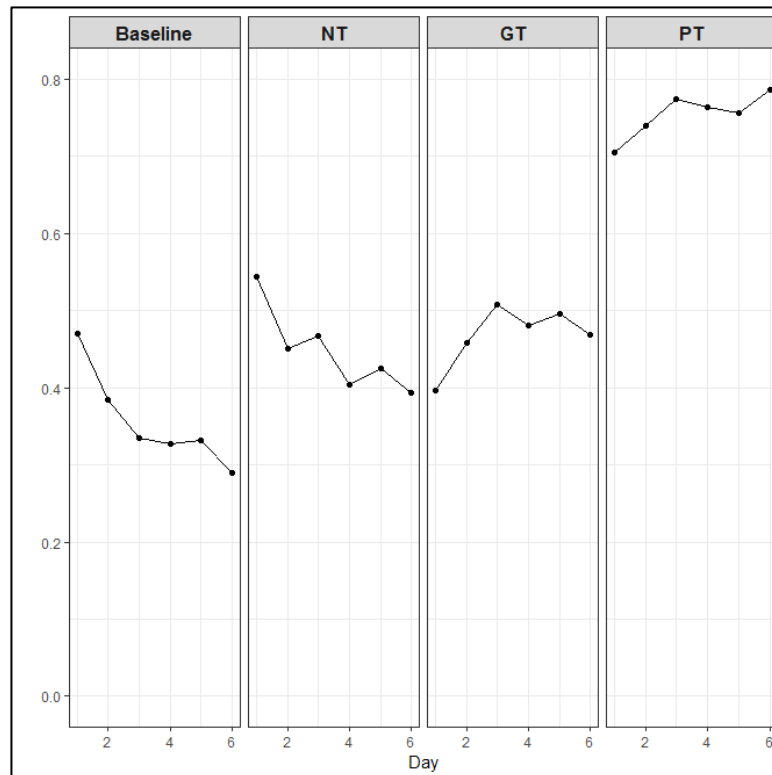


Figure 2 also reveals that in *PT* the percentage of healthy choices keeps increasing, reaching almost 80% of “good” items chosen on the last day of the three-week intervention. So, it seems that, on top of the potential competition created by the grades, keeping parents informed about their children’s “performance” further affected participants’ behavior. However, in *Baseline* and *NT*, subjects do not have any reference point for competition. This fact, combined with the lack of incentives for choosing the healthy option, leads to a reduction in the proportion of healthy items chosen over time in these treatments.

To study the extent to which children react by picking healthier items solely due to the knowledge that their average grade will be released to their parents, we look at students’ decisions in the first period in each treatment. Because all

contributions matched led to a substantial increase in charitable contributions. Majolo and Marechal (2017) find greater within-group cooperation when groups of children were competing with other groups.

parents received exactly the same information regarding the experiment in all treatments, children's behavior in period 1 in *PT* is not affected by their parents' direct intervention but rather only by the expectation of their information to be released to their parents in the future.

As Figure 2 shows, the percentage of healthy choices in the first period is 70.52% in *PT*, which is larger than in *Baseline* (46.68%), *GT* (39.68%), and *NT* (54.37%).⁶³ These results indicate that simply knowing that their parents will view this information has a large effect on the food choices made by the students.

3.5.3 Longer-term effects after the intervention period and the removal of incentives

A key issue is whether the effects of the intervention persist over time. While some previous work has found that there are strong effects from paying children to choose the healthier option, there is little evidence that the effects persist much after the payments cease. A mechanism that produces enduring effects is eminently more practical than one that does not, particularly if cash payments are not required indefinitely. This section analyzes the effect on children's decisions several months after the incentives are removed. As explained in the experimental design, four months after the end of the intervention and within the next academic year, we ran a surprise session. In this case, all subjects participated in a one-shot *Baseline*, so the previous incentives were not in play.

⁶³ Since in the first period all subjects in a given school are independent, we conduct cluster-adjusted chi-square tests for individual-level data. Differences are statistically significant for the comparison *PT* versus *Baseline* ($\chi^2 = 0.445$, $p = 0.063$), for *PT* versus *GT* ($\chi^2 = 4.140$, $p = 0.042$), and for *PT* versus *NT* ($\chi^2 = 2.746$, $p = 0.097$). Cluster-adjusted, one-tailed t -tests give $t = 3.027$, $p = 0.019$ versus *Baseline*, $t = 4.089$, $p = 0.007$ versus *GT*; and $t = 2.162$, $p = 0.048$ versus *NT*. Tests with school-level data also indicate significance at the 5% level.

Table 2 presents a summary similar to Table 1, including the average individual grades and the proportion of healthy choices made by individuals in the surprise session.⁶⁴

Table 2. Average grades and healthy food choices in the surprise session, by treatment

Initial Treatment	Observations	Subjects	Average grade	% healthy choices
<i>Baseline</i>	33	33	5.06 (1.70)	0.41 (0.25)
<i>NT</i>	68	68	5.60 (1.77)	0.47 (0.26)
<i>GT</i>	65	65	5.94 (1.91)	0.53 (0.27)
<i>PT</i>	45	45	7.43 (1.65)	0.69 (0.21)

Notes: Standard deviations are reported in parentheses.

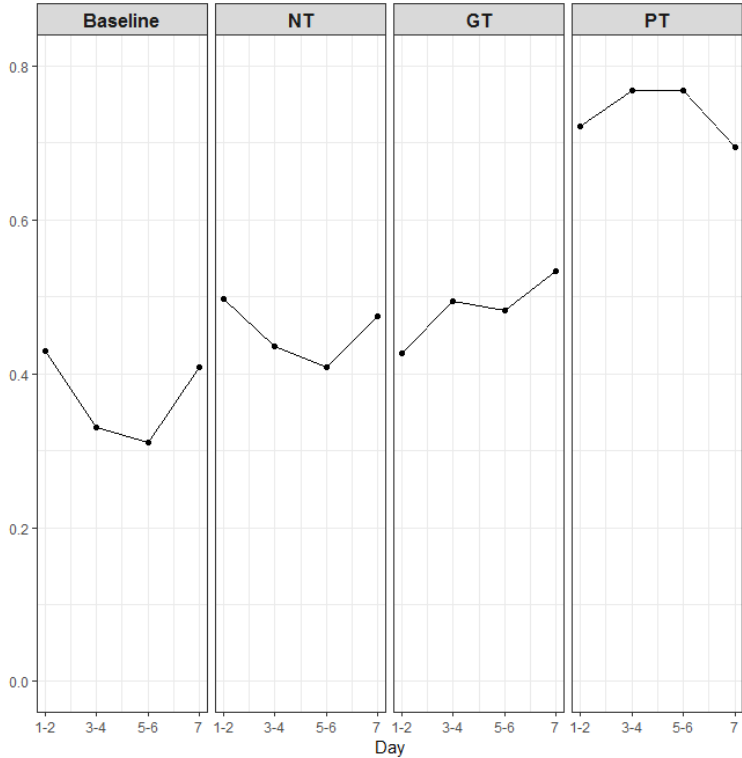
Even after removing the incentives and four months after the intervention period, the proportion of healthy choices in each of *GT* (53%), *NT* (47%) and *PT* (69%) are larger than in *Baseline* (41%). A test of proportions at the individual level shows that the differences are significant for *GT* and *PT* versus *Baseline* ($p = 0.025$, and $p = 0.000$, respectively), but not when we compare *NT* and *Baseline* ($p = 0.259$). Comparing choices in *PT* to *NT* and *GT*, we find highly significant differences ($p = 0.000$ for one-tailed tests of proportions for the comparisons *PT*

⁶⁴ The lower number of observations in the surprise sessions is mainly due to the fact that one school in *Baseline* and one school in *PT* decided not to participate in the surprise sessions.

vs. *GT* and *PT* vs. *NT*).⁶⁵ These results indicate two main conclusions. First, there is a positive effect of *GT* on healthy choices while the effect of the incentive weakens over time in *NT*. Second, providing information to the parents has a very strong effect, which persists even four months after the incentive was removed. This leads to dramatically healthier choices.

We next focus on the comparison of subjects' decisions in the surprise session to their average behavior in the first part (first six days) of the experiment. Figure 3 compares the average percentage of healthy choices in the first stage and the surprise session.

Figure 3. Percentage of healthy choices on days 1-6 and on day 7 (surprise session)



We see that children did not behave differently on day 7 (the surprise session) compared to their average behavior during the first stage (i.e., days 1-6). In *Baseline*, the percentage of healthy choices improves from 36% in the first

⁶⁵ The results are similar if we take the average choices at the school level and conduct one-tailed Mann-Whitney tests or *t*-tests (with *p*-values ranging from 0.02 to 0.06).

stage of the experiment to 41% in the surprise session; in GT this improves from 46% to 53%; in NT it improves from 45% to 47%; and, finally, in PT the percentage drops from 74% to 69%. Differences between the average of the first six days and the surprise session are not statistically significant in any treatment even with one-tailed Wilcoxon signed-rank tests ($Z = -1.342$, $p = 0.179$; $Z = -1.342$, $p = 0.179$; $Z = -1.069$, $p = 0.285$; and $Z = 1.342$, $p = 0.179$, for *Baseline*, *GT*, *NT*, and *PT*, respectively).⁶⁶ So, it appears there is a carry-over effect for all treatments, which would lead to healthier choices in the cases in which the incentives were effective in the first place.

Result 3: *Participants' behavior in the "surprise session" of the experiment was quite similar to their average behavior in the three weeks intervention period when the different incentives were in place, showing a remarkable degree of persistence. This indicates significantly healthier behavior even four months after the incentives have been removed.*

3.5.4 Regression analysis

To better understand the mechanisms underlying the group-level patterns reported above, Table 3 reports the results from linear regression models in which the dependent variable is subject i 's grade in period t . All specifications use cluster-robust standard errors at the school level to account for heteroskedasticity and intra-school correlation.⁶⁷

⁶⁶ Conclusions remain the same when we look at the average grades, also when using t -tests and permutation tests.

⁶⁷ Since all the explanatory variables included in the models presented in Table 3 are time-invariant, these specifications do not include individual, school or period (day) fixed effects. For robustness, Table E.5 in the Appendix reports estimates from multi-level models including random effects to account for time-invariant individual heterogeneity, within-school correlation and common temporal shocks affecting all subjects. Multi-level models also provide an alternative tool for dealing with clustered data (e.g., Primo *et al.*, 2007), and Bayesian inferential methods yield accurate estimates for such hierarchical models even with as few as 3 clusters (Gelman, 2006). The main results are similar to those presented in Table 3.

However, since the number of schools in our sample is quite small (12), conventional inference methods may be unreliable because large-sample assumptions do not hold, and standard errors may be biased downwards (Donald and Lang, 2007; Cameron *et al.*, 2008). Hence, the table also reports p -values for the two-sided Wald hypothesis tests computed according to Cameron *et al.*'s (2008) wild bootstrap- t procedure for clustered data. This procedure is useful for small sample sizes like ours because it does not rely on asymptotic approximations and has been shown to perform well with very small numbers of clusters (Cameron *et al.* 2008; MacKinnon and Webb 2017; Djogbenou *et al.* 2018).

Table 3. Regression analysis of subjects' average grades

Table 3. Regression analysis of subjects' average grades

	Data from intervention period			Four months after intervention
	(1)	(2)	(3)	(4)
<i>Constant</i>	4.72*** (0.05) [0.00]	6.01*** (0.30) [0.00]	4.80*** (0.49) [0.00]	5.04*** (0.11) [0.00]
<i>NT</i>	0.74* (0.24) [0.07]	0.95* (0.27) [0.07]	0.89 (0.43) [0.12]	0.56** (0.17) [0.04]
	1.04*** (0.30) [0.00]	1.23** (0.27) [0.03]	1.68** (0.34) [0.02]	0.90* (0.30) [0.06]
<i>PT</i>	3.15*** (0.21) [0.00]	3.11** (0.17) [0.02]	3.04*** (0.38) [0.00]	2.39*** (0.25) [0.00]
<i>Male</i>		0.31 (0.18) [0.11]	0.49** (0.138) [0.03]	
<i>Average school grade</i>		0.141 (0.07) [0.07]	0.08 (0.09) [0.40]	
<i>Personal Impulsivity</i>		-0.081* (0.41) [0.09]	-1.19** (0.47) [0.04]	
<i>School Impulsivity</i>		-0.35* (0.15) [0.06]	-0.23 (0.20) [0.31]	
<i>Parents hold University degree</i>			-0.11 (0.24) [0.69]	
<i>Household Income</i>			0.62 (0.25) [0.24]	
<i>#Observations</i>	1378	1351	945	213
<i>R²</i>	0.30	0.33	0.39	0.17

Notes: Estimation method is OLS. Cluster-robust standard errors clustered by school in parentheses (first line below the coefficients). The *p*-values for two-sided Wald tests computed according to Cameron *et al.*'s (2008) wild bootstrap-*t* procedure accounting for small number of clusters (schools) are reported in brackets (second line below the coefficients). Significance levels (based on the – more conservative, bootstrapped – Wald tests): ***, **, and * denote significance at *p* = 0.01, 0.05, and 0.10, respectively.

The explanatory variables in the first column of Table 3 are the indicators for each treatment (*NT*, *GT*, and *PT*, taking *Baseline* as the benchmark). Subjects' average grade – associated with their choice of healthy food items – increases in all three treatments compared to *Baseline*, even after accounting for the small number of schools in our sample. This result suggests that participants generally responded to incentives, and reinforces the evidence reported in Table 1. Moreover, and also in line with the previous results, we observe that subjects' average grade is significantly larger in *PT* than in *GT* ($F = 39.84$, $p = 0.000$; test for equality of coefficients) and *NT* ($F = 75.37$, $p = 0.000$). There is no statistically significant difference between the effect of the grades-only treatment and the effect of the nutritionist, though ($F = 0.63$, $p = 0.45$; test for equality of coefficients between *GT* and *NT*).⁶⁸

Column (2) incorporates additional controls to the covariates of the first column and shows that the treatment effects remain robust to controlling for subject specific characteristics. Column (3) replicates the analysis in column (2) but takes also into account the socio-economic level of a child's family. Specifically, *Parents hold University degree* is a binary variable that takes value 1 if at least one of the parents achieved graduate or postgraduate education and 0 otherwise; and *Household Income* is a composite index of the number of durable goods (laptop and desktop computer, tablet, electronic books, cars) available in each student's home.^{69,70} We observe that a higher educational level

⁶⁸ Table E.6 in the Appendix reports estimates from probit models in which the outcome is a binary variable taking the value 1 if subject i chose a healthy food item (grade 7.5 or above) in period t . We also present p -values for Wald hypothesis tests computed according to Kline and Santos' (2012) score bootstrap procedure for clustered data, which adapts the wild bootstrap to maximum likelihood estimators and has been shown to perform remarkably well even with very few clustering units. The substantive results are similar to those in Table 3.

⁶⁹ Using consumption-based measures of household welfare or resources – instead of relying on self-reported income – is a common practice in view of the fact that income is more likely to be mis-measured and under-reported (e.g., Meyer and Sullivan, 2011).

⁷⁰ A potential source of concern regarding the specifications in columns (2) and (3) is that the variables measuring students' impulsivity and households' socio-economic characteristics were obtained from the

for parents does not significantly affect children's grade, and neither does the household's economic level once we account for the small number of sample schools through the use of the wild bootstrap.⁷¹ Compared to the specification in column (2), the coefficient for *Average school grade* (as a measure of academic performance) ceases to be statistically significant, while that for *Male* now becomes significant. As for the treatment effects, subjects' average grades (from food choice) remain systematically higher under *PT*, *GT* and *NT* than under the *Baseline* treatment in this specification, although the effect of the nutritionist becomes marginally insignificant once we account for small cluster size.

Finally, Column (4) repeats the specification in column (1) for the data in the surprise session. Note that all subjects participated in the *Baseline* in the surprise session, so the treatment dummies in column (4) capture the persistence from the original treatments. We report that the positive effect found for *NT*, *GT* and *PT* still remains in the long run.

In order to assess the evolution of the treatment effects over time, Figure 4 plots the differences in students' average marks between the *Baseline* treatment and *GT*, *NT* and *PT* for each period (i.e., the daily treatment effects).⁷² Results show that differences between *PT* and the *Baseline* are positive and statistically significant throughout the experiment, and much larger than what we

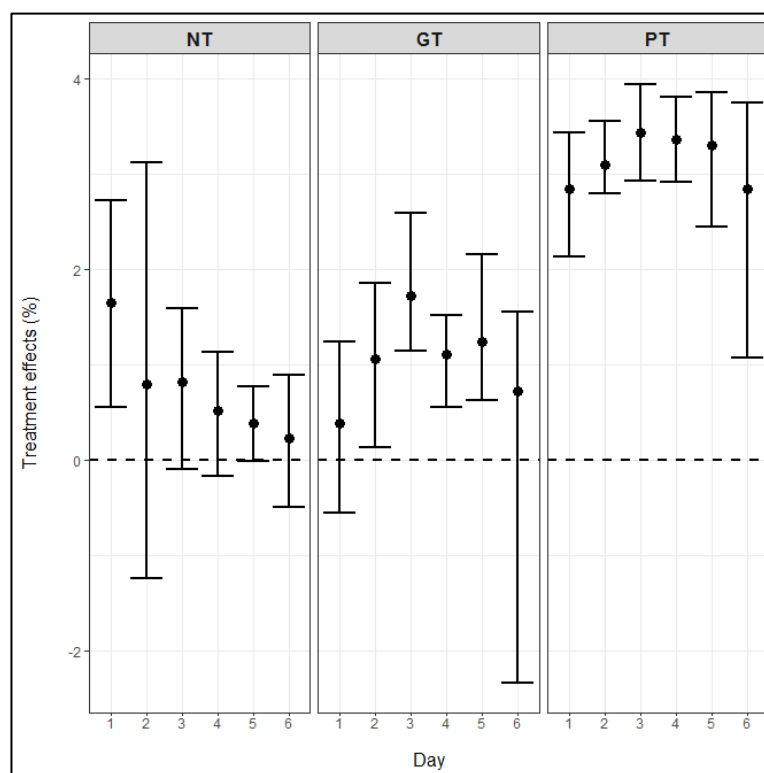
response of subjects, teachers, and parents to surveys conducted post-treatment. While it is highly unlikely that the treatments affected subjects' impulsivity or fixed family characteristics, it is in principle conceivable that the survey responses (or unobservables related to these characteristics) could have been influenced by the experimental intervention. That said, the fact that the treatment effect estimates are significant in column (1) – see also Tables E.5 and E.6 in the Appendix – and are robust across specifications mitigates this concern.

⁷¹ It is true that children from lower socioeconomic background have been shown (e.g., Muller *et al.*, 2005) to be much less responsive to interventions aimed at tackling obesity and overweight. However, note that our population is quite homogeneous, so a lower socioeconomic background would still be a middle class family (medium-low). In the same way, a high-level family would be medium-high. We do not have extreme cases in our population.

⁷² The per-period effects in Figure 4 are based on the parameter estimates reported in Table E.7 in the Appendix, which – like the specifications in columns (1) and (4) of Table 3 – include only the treatment indicators as predictors. The confidence intervals in Figure 4 are obtained by inverting the wild bootstrap-*t* tests, and are not necessarily symmetric (Roodman *et al.*, 2019). They are typically more conservative than “standard” confidence sets relying on asymptotic approximations.

see in the panels referring to treatments *GT* and *NT*. The differences between *GT* and the *Baseline* become significant from the second period onwards, and remain at least two standard deviations greater than zero until the last day (6) of the intervention period. For *NT*, the average grade is significantly higher than in the *Baseline* treatment only on the first day (when the nutritionist gives the talk). These findings support the idea that the positive effect of the treatment dummies – especially *PT* and, to a lesser extent, *GT* – on students' healthy food choices comes from a stable effect that affects decisions in every period rather than being just a strong initial effect that vanishes over time.

Figure 4. Daily differences in subjects' grades between each treatment (*NT*, *GT*, and *PT*) and the *Baseline* during the intervention period



Notes: For each period (day), the figure plots the difference in average subject's grades between *NT* (left panel), *GT* (middle panel) and *PT* (right panel) and the *Baseline* treatment, in percentage points. Solid circles represent point estimates; vertical lines give the 95% confidence intervals, obtained by inverting the wild bootstrap-*t* tests.

3.6 Conclusions

A critical contemporary problem is that of poor diet and obesity. This has been linked to a variety of health problems, with concomitant economic consequences. Perhaps this is most important for children, where there are long-lasting health effects; addressing this issue at the root should offer the best hope for the future. Data from a variety of sources indicate that children's eating patterns are far from encouraging, since children still fail to meet recommendations for the daily consumption of fruit and vegetables. Childhood obesity rates in the U.S. tripled from 1971-1974 and 2011-2012.

We conduct a field intervention designed to test whether it is possible to beneficially affect the diets of children. Previous work has shown positive effects with contemporaneous benefits, but at best limited enduring effects thereafter. Our design shows that non-material incentivizes can be effective for leading children to make healthier food choices at school. We align children's appraisal of food choices with their appraisal of schoolwork by introducing a system in which food items are graded based on their nutritional value.

Critically, we also involve parents as change agents, providing them with information regarding the food choices of their children. While providing information about grades and advice from nutritionists have definite value (and the effect from providing information about grades seems to persist), involving the parents in the decision process generates by far the biggest boost in healthy eating. This provides us with very strong results *that are undiminished four months after our intervention was completed*.

We feel that our results are extremely promising, since our second-generation approach involves little or no financial cost and requires only monitoring from parents or peers. It is obvious that more research would be

helpful, since our study is limited to middle-class families in Spain. If policymakers were able to establish good dietary habits in early childhood, this would make great inroads on the current set of health problems resulting from poor nutritional habits and obesity.

4. Bibliography

- Aguiar, F, Branas-Garza, P., Cobo-Reyes, R., Jimenez, N., and Miller, L. (2009) “Are women expected to be more generous?” *Experimental Economics*, 12: 93-98.
- Aigner, D. and Cain, G. (1977) “Statistical Theories of Discrimination in Labor Markets” *Industrial and Labor Relations Review*, 30(2): 175-187
- Akerlof, G. and Kranton, R. (2000) “Economics and Identity” *Quarterly Journal of Economics*, 115: 715-753.
- Arcidiacono, P., Hotz, J., and Kang, S. (2012) “Modeling College Major Choice using Elicited Measures of Expectations and Counterfactuals.” *Journal of Econometrics*, 166 (1): 3-16.
- Babcock, P., Bedard, K., Charness, G., Hartman, J., and Royer, H. (2015) “Letting down the team? Social effects of team incentives.” *Journal of the European Economic Association*, 13(5): 841-870
- Belot, M., James, J., and Nolen, P. (2016) “Incentives and children's dietary choices: A field experiment in primary schools,” *Journal of Health Economics*, 50: 213-229.
- Belot, M., Berlin, N., James, J., and Skafida, V. (2019) “The formation and malleability of dietary habits: A field experiment with low income families,” Mimeo.
- Billig, M. and Tajfel, H. (1973), “Social Categorization and Similarity in Intergroup Behaviour,” *European Journal of Social Psychology*, 3: 27-52.
- Black, S., and Strahan, P. (2001) “The Division of Spoils: Rent-Sharing and Discrimination in a Regulated Industry,” *American Economic Review*, 91: 814–831.
- Bohnet, I., and Frey, B. (1999) “Social Distance and other-regarding Behavior in Dictator Games: Comment” *The American Economic Review*, 89(1): 335-339
- Bohnet, I., van Geen, A., and Bazerman, M. (2016) “When Performance Trumps Gender Bias: Joint versus Separate Evaluation,” *Management Science*, 5: 1225–1234.
- Bordalo, P., Coffman, K., Gennaioli, N. and Shleifer, A. (2016) “Stereotypes” *The Quarterly Journal of Economics*: 1753–1794
- Bordalo, P., Coffman, K., Gennaioli, N. and Shleifer, A. (2016a) “Beliefs about Gender” NBER working paper.

- Borgmeier, I., and Westenhoefer, J. (2009) "Impact of different food label formats on healthiness evaluation and food choices of consumers: a randomized-controlled study," *BMC Public Health*, 9: 184.
- Bornstein, G, Gneezy, U., and Nagel, R. (2002) "The effect of intergroup competition on group coordination: an experimental study," *Games and Economic Behavior*, 41(1): 1-25.
- Branas-Garza, P. (2006) "Poverty in Dictator Games: Awakening Solidarity." *Journal of Economic Behavior and Organization*, 60: 306-320.
- Branas-Garza, P. (2007) "Promoting helping behavior with framing in Dictator Games." *Journal of Economic Psychology*, 28: 477-486.
- Buser, T. Niederle, M., and Oosterbeek, H. (2014) "Gender, Competitiveness and Career Choices." *Quarterly Journal of Economics*, 129 (3): 1409-1447.
- Cameron, A.C., Gelbach, J.B., and Miller, D.L. (2008) "Bootstrap-base improvements for inference with clustered standard errors," *Review of Economics and Statistics*, 90(3): 414-427.
- Charness, G., Cobo-Reyes, R. and Jimenez, N. (2014) "Identities, Selection, and Contributions in a Public-goods Game." *Games and Economic Behavior*, 87: 322-338
- Charness, G. and Gneezy, U. (2008) "What's in a name? Anonymity and social distance in dictator and ultimatum games." *Journal of Economic Behavior and Organization*, 68: 29-35
- Charness, G., and Gneezy, U. (2009) "Incentives to exercise," *Econometrica*, 77(3): 909-931.
- Charness, G. and Holder, P. (2018) "Charity in the Laboratory: Matching, Competition, and Group Membership" forthcoming at *Management Science*.
- Charness, G., Masclet, D., and Villeval M.C. (2014) "The dark side of competition for status," *Management Science*, 60(1): 38-55.
- Charness, G., Rigotti, L., and Rustichini, A. (2007), "Individual Behavior and Group Membership," *American Economic Review*, 97, 1340-1352.
- Chen, Y. and S. Li (2009), "Group Identity and Social Preferences," *American Economic Review*, 99(1), 431-57.
- Coate, S. and Loury, G. (1993) "Will affirmative-action policies eliminate negative stereotypes?" *American Economic Review*, 83(5): 1220-1240

- Coffman, Katherine B. (2014) “Evidence on self-stereotyping and the contribution of ideas” *Quarterly Journal of Economics*, 129(4): 1625-1660.
- Cornell, B. and Welch, I. (1996) “Culture, Information, and Screening Discrimination” *Journal of Political Economy*, 104(3): 542-571
- Correll, S.J. (2001). “Gender and the career choice process: The role of biased self-assessments.” *American Journal of Sociology* 106(6):1691–1730.
- Croson, R. and Gneezy, U. (2009) “Gender Differences in Preferences.” *Journal of Economic Literature*, 47: 448-474.
- Croson, R., and Shang, J. (2008), “The Impact of Downward Social Information on Contribution Decisions” *Experimental Economics*, 11: 221-233
- Dale, D. (2004), “Charitable Lottery Structure and Fund Raising: Theory and Evidence” *Experimental Economics*, 7: 217-234
- Deci, E. L., & Ryan, R. M. (1985) *Intrinsic motivation and self-determination in human behavior*. New York: Plenum.
- DeCosta, P., Moller, P., Frost, M.B., and Olsen, A. (2017) “Changing children’s eating behaviour – A review of experimental research,” *Appetite*, 113: 328-357.
- Djogbenou, A.A., MacKinnon, J.G., and Nielsen, M.Ø. (2018). “Asymptotic theory and wild bootstrap inference with clustered standard errors,” QED Working Paper 1399, Department of Economics, Queen’s University.
- Donald, S. and Lang, K. (2007). “Inference with difference-in-differences and other panel data,” *Review of Economics and Statistics*, 89(2): 221–233.
- Donner, A., and Klar, N. (2000). *Design and Analysis of Cluster Randomization Trials in Health Research*. London: Arnold.
- Downs, J. S., Wisdom, J., Wansink, B., and Loewenstein, G. (2013) “Supplementing menu labeling with calorie recommendations to test for facilitation effects,” *American Journal of Public Health*, 103(9): 1604-1609.
- Eckel, C. and Grossman, P. (1996) “Altruism in Anonymous Dictator Games.” *Games and Economic Behavior*, 16: 181-191.
- Eckel, C and P. Grossman (2008), “Subsidizing Charitable Contributions: A Natural Experiment Comparing Matching and Rebate Subsidies,” *Experimental Economics*, 11: 234-252

- Edwards, J. and List, J. (2014) "Toward an Understanding of why Suggestions Work in Charitable Fundraising: Theory and Evidence from a Natural Field Experiment." *Journal of Public Economics*, 114: 1-13.
- Ellison, G., and Swanson, A. (2010) "The Gender Gap in Secondary School Mathematics at High Achievement Levels: Evidence from the American Mathematics Competitions," *Journal of Economic Perspectives*, 24(2): 109-128.
- Esarey, J., and Menger, A. (2019) "Practical and effective approaches to dealing with clustered data," *Political Science Research and Methods*, 7(3): 541-599.
- Fershtman, C., Gneezy, U., and List, J. (2012) "Equity aversion: Social norms and the desire to be ahead," *American Economic Journal: Microeconomics*, 4(4): 131-44.
- Fiedler, M. and Haruvy, E. (2009) "The lab versus the virtual lab and virtual field: an experimental investigation of trust games with communication." *Journal of Economic Behavior & Organization*: 72(2), 716–724.
- Fiedler, M., Haruvy, E., and Li, S. (2011) "Social distance in a virtual world experiment." *Games and Economic Behavior*: 72(2), 400–426.
- Fischbacher, U. (2007). "Z-Tree: Zurich toolbox for ready-made economic experiments." *Experimental Economics*, 10: 171-178.
- Fisher, J. and Birch, L. (1999) "Restricting access to palatable foods affects children's behavioral response, food selection, and intake," *American Journal of Clinical Nutrition*, 69(6): 1264-1272.
- Frey, B. and Meier, S. (2004), "Social Comparisons and Pro-social Behavior: Testing "Conditional Cooperation" in a Field Experiment," *American Economic Review*, 94(5): 1717- 1722.
- Galloway, A. T., Fiorito, L. M., Francis, L. A., and Birch, L. L. (2006) "Finish your soup: Counterproductive effects of pressuring children to eat on intake and affect," *Appetite*, 46(3): 318-323.
- Gelman, A. (2006) "Prior distributions for variance parameters in hierarchical models," *Bayesian Analysis*, 1(3):515–33.
- Gerber, A., Karlan, D., and Bergan, D. (2009) "Does the media matter? A field experiment measuring the effect of newspapers on voting behavior and political opinions," *American Economic Journal: Applied Economics*, 1(2): 35-52.

- Gigerenzer, G., Galesic, M., & Garcia-Retamero, R. (2013). "Stereotypes About Men's and Women's Intuitions: A Study of Two Nations." *Journal of Cross-Cultural Psychology*, 45: 62-81.
- Global Burden of Disease (2016). *The Lancet*. Available at <http://www.thelancet.com/gbd>, accessed 02 January 2018.
- Goldin, C., and Rouse, C. (2000) "Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians," *American Economic Review*, XL, 715–742.
- Greiner, B. (2015). "Subject pool recruitment procedures: organizing experiments with ORSEE." *Journal of the Economic Science Association*, 1(1): 114-125
- Guiso, L., Monte, F., Sapienza, P., and Zingales, L. (2008). "Culture, math, and gender." *Science* 320(5880): 1164–1165.
- Günther, C., Ekinici, N.A., Schwierén, C., Strobel, M., (2010) "Women can't jump?—An experiment on competitive attitudes and stereotype threat", *Journal of Economic Behavior and Organization*, 75 (3): 395
- Haire-Joshu, D., and Tabak. R. (2016) "Preventing obesity across generations: Evidence for early life intervention," *American Review of Public Health*, 37: 253-271.
- Halladay, B. (2017). "Perception Matters: The Role of Task Gender Stereotype on Confidence and Tournament Selection" UCSB working paper.
- Hansen, B.B., and Bowers, J. (2008) "Covariate balance in simple, stratified and cluster comparative studies," *Statistical Science*, 23(2): 919-936.
- Hedges, L., and Nowel, A. (1995) "Sex differences in mental test scores, variability, and number of High-Scoring individuals" *Science*, 269: 41-45
- Hill, C., Corbett, C., and St. Rose, A. (2010). Why So Few? Women in Science, Technology, Engineering, and Mathematics. American Association of University Women: Washington, DC.
- Hoffman, E., McCabe, K., and Smith, V. (1996) "Social Distance and Other-Regarding Behavior in Dictator Games." *The American Economic Review*, 86: 653-660
- Huberman, B., Loch, C. and A. Öncüler (2004) "Status as a valued resource," *Social Psychology Quarterly*, 67(1): 103-114.
- Huck, S. and Rasul, I. (2011), "Matched Fundraising: Evidence from a Natural Field Experiment," *Journal of Public Economics*, 95: 351-362.
- Hyde, J.S., and Mertz, J.E. (2009). Gender, culture, and mathematics performance. *Proceedings of the National Academy of Sciences* 106(22):8801–8807.

- Hyde, J.S., Lindberg, S.M., Linn, M.C., Ellis, A.B., and Williams, C.C. (2008). Diversity. Gender similarities characterize math performance. *Science* 321(5888):494–495.
- Ibragimov, R, and Muller, U.K. (2010) “t-Statistic based correlation and heterogeneity robust inference,” *Journal of Business & Economic Statistics*, 28(4): 453-468.
- Jaime, P.C., and Lock, K. (2009) “Do school based food and nutrition policies improve diet and reduce obesity?” *Preventive Medicine*, 48(1): 45-53.
- Jansen, E., Mulken, S., and Jansen, A. (2007) “Do not eat the red food!: Prohibition of snacks leads to their relatively higher consumption in children,” *Appetite*, 49(3): 572-577.
- Jansen, E., Mulken, S., Emond, Y., and Jansen, A. (2008) “From the garden of Eden to the land of plenty. Fruit and sweets intake leads to increased fruit and sweets consumption in children,” *Appetite*, 51(3): 570-575.
- Janusonis, S. (2009) “Comparing two small samples with an unstable, treatment-independent baseline,” *Journal of Neuroscience Methods*, 179(2): 173-178.
- Joensen, J. S. and Nielsen, H. S. (2016), “Mathematics and Gender: Heterogeneity in Causes and Consequences.” *Economic Journal* 126: 1129-1163.
- Just, D.R., and Price, J. (2013) “Using incentives to encourage healthy eating in children,” *Journal of Human Resources*, 48(4): 855-872.
- Karlan, D. and List, J. (2007), “Does Price Matter in Charitable Giving? Evidence from a Large-Scale Natural Field Experiment,” *American Economic Review*, 97: 1774-1793.
- Karlan, D., List, J., and Shafir, E. (2011), “Small matched and Charitable giving: Evidence from a natural field experiment” *Journal of Public Economics*, 95: 344-350.
- Kiefer, A.K., and Sekaquaptewa, D. (2007). “Implicit stereotypes, gender identification, and math- related outcomes: A prospective study of female college students.” *Psychological Science* 18(1): 13–18.
- Kim S., Moore L., and Galuska D. (2014) “Vital signs: fruit and vegetable intake among children – United States, 2003–2010,” *Morbidity and Mortality Weekly Report*, 63(31): 671– 676.
- Kline, P., and Santos, A. (2012) “A score based approach to wild bootstrap inference,” *Journal of Econometric Methods*, 1: 23-41.

- Landry, C., Lange, A., List, J., Price, M., and Rupp, N. (2006) "Toward An Understanding of The Economics of Charity: Evidence From a Field Experiment," *Quarterly Journal of Economics*, 121: 747-782.
- Lichtenstein, S., Fischhoff, B., and Phillips, L. (1982) "Calibration of Probabilities: The State of the Art to 1980." In *Judgment under Uncertainty: Heuristics and Biases*, ed. Daniel Kahneman, Paul Slovic, and Amos Tversky, 306–34. Cambridge and New York: Cambridge University Press.
- Lim, C.U. and Harrell, D.F. (2015) "Understanding players' identities and behavioral archetypes from avatar customization data." *Proceedings of the IEEE Computational Intelligence and Games*: 238-245.
- List, J., and Lucking-Reiley, D. (2002), "The Effects of Seed Money and Refunds on Charitable Giving: Experimental Evidence from a University Capital Campaign." *Journal of Political Economy* 110: 215-233.
- List, J.A., and Samek, A.S. (2015) "The behavioralist as nutritionist: Leveraging behavioral economics to improve child food choice and consumption," *Journal of Health Economics*, 39: 135-146.
- Loewenstein, G., Price, J., and Volpp, K. (2016) "Habit formation in children: Evidence from incentives for healthy eating," *Journal of Health Economics*, 45: 47-54.
- Loewenstein, G., Weber, E., Hsee, C., and Welch, N. (2001) "Risk as Feelings." *Psychological Bulletin*, 127(2): 267–86.
- Ludbrook, J., and Dudley, H. (1998) "Why permutation tests are superior to t and F tests in biomedical research," *The American Statistician*, 52(2): 127-132.
- Lundberg, S. and Startz, R. (1983) "Private discrimination and social intervention in competitive labor market" *American Economic Review*, 73(3): 340-347
- Lundeberg, M., Fox, P. and Punccohar, J. (1994) "Highly Confident but Wrong: Gender Differences and Similarities in Confidence Judgements." *Journal of Educational Psychology*, 86(1): 114–21.
- Lynch, C., Kristjansdottir, A., Te Velde, S., Lien, N., Roos, E., and Thorsdottir, I. (2014) "Fruit and vegetable consumption in a sample of 11-year-old children in ten European countries – the PRO GREENS cross-sectional survey," *Public Health Nutrition*, 17(11): 2436-2444.
- MacKinnon, J.G., and Webb, M.D. (2017) "Pitfalls when estimating treatment effects with clustered data," *The Political Methodologist*, 24(2): 20-31.

- Majolo, B., and Marechal, L. (2017) "Between-group competition elicits within-group cooperation in children," *Nature Scientific Reports*, 7: 43277.
- Martin, R. and Randal, J. (2008), "How is Donation Behaviour Affected by the Donations of Others?" *Journal of Economic Behavior and Organization*, 67: 228-238.
- Meyer, B. and Sullivan, J. (2011) "Further results on measuring the well-being of the poor using income and consumption," *Canadian Journal of Economics*, 44(1): 52-87.
- Micha R, Peñalvo JL, Cudhea F, Imamura F, Rehm CD, and Mozaffarian D. (2017) "Association between dietary factors and mortality from heart disease, stroke, and Type 2 diabetes in the United States," *Journal of the American Medical Association*, 317(9): 912–924.
- Moffitt, T., Arseneault, L., Belsky, D., Dickson, N., Hancox, R., Harrington, H., Houts, R., Poulton, R., Roberts, B., Ross, Malcolm, B., Sears, R., W., and Caspi, A. (2011) "A gradient of childhood self-control predicts health, wealth, and public safety," *Proceedings of the National Academy of Sciences*, 108 (7): 2693-2698.
- Morgan, J., and Sefton, M. (2000), "Funding Public Goods with Lotteries: Experimental Evidence," *Review of Economic Studies*, 67: 785-810
- Morizet, D., Depezay, L., Combris, P., Picard, D., and Giboreau, A. (2012) "Effect of labeling on new vegetable dish acceptance in preadolescent children," *Appetite*, 59: 399-402.
- Niederle, M. Segal, C., and Vesterlund, L. (2013), "How Costly is Diversity? Affirmative Action in Light of Gender Differences in Competitiveness," *Management Science*, 59(1), 1-16.
- Niederle, M., and Vesterlund, L. (2007). "Do women shy away from competition? Do men compete too much?" *Quarterly Journal of Economics* 122(3):1067–1101.
- Niederle, M., and Vesterlund, L. (2011) "Gender and Competition", *Annual Review in Economics*, 3: 601–30.
- Noble, E., and Kanoski, S. (2016) "Early life exposure to obesogenic diets and learning and memory dysfunction," *Current Opinion in Behavioral Sciences*, 9: 7-14.
- Ogden, J., Cordey, P., Cutler, L., and Thomas, H. (2013) "Parental restriction and children's diets. The chocolate coin and Easter egg experiments," *Appetite*, 61: 36-44.

- Patall, E.A., Cooper, H., and Robinson, J.C. (2008) "The effects of choice on intrinsic motivation and related outcomes: A meta-analysis of research findings," *Psychological Bulletin*, 134(2): 270-300.
- Pelchat, M., and Pliner, P. (1995) "Try it. You'll like it. Effects of information on willingness to try novel foods," *Appetite*, 24(2): 153-165.
- Pesarin, F., and Salmaso, L. (2010) *Permutation Tests for Complex Data. Theory, Applications and Software*. Chichester: Wiley.
- Phelps, E. (1972) "The Statistical Theory of Racism and Sexism" *American Economic Review*, 62(4): 659-661
- Pinkston, J. (2006) "A Test of Screening Discrimination with Employer Learning" *Industrial and Labor Relations Review*, 59(2): 267-284
- Primo, D.M., Jacobsmeier, M.L., and Milyo, J. (2007) "Estimating the impact of state policies and institutions with mixed-level data," *State Politics and Policy Quarterly*, 7(4): 446-459.
- Putnam, R. (2000), *Bowling Alone: The Collapse and Revival of American Community*, Simon and Shuster: New York.
- Raab, G.M., and Butcher, I. (2005) "Randomization inference for balanced cluster-randomized trials," *Clinical Trials*, 2(2): 130-140.
- Reuben, E., Sapienza, P., and Zingales, L. (2014) "How stereotypes impair women's career in science." *Proceedings of the National Academy of Sciences*, 111: 4403-4408.
- Ridder, M.A., Heuvelmans, M.A., Visscher, T.L, Seidell, J.C., and Renders, C.M. (2010) "We are healthy so we can behave unhealthily: A qualitative study of the health behaviour of Dutch lower vocational students," *Health Education*, 110(1): 30-42.
- Rigal, N., Rubio, B., and Monnery-Patris, S. (2016) "Is harsh caregiving effective in toddlers with low inhibitory control? An experimental study in the food domain," *Infant Behavior and Development*, 43: 5-12.
- Roberto, C., Larsen, P., Agnew, H., Baik, J., and Brownell, K. (2010) "Evaluating the impact of menu labelling on food choices and intake," *American Journal of Public Health*, 100(2): 312-318.
- Roodman, D., Nielsen, M.Ø., MacKinnon, J.G., and Webb M.D. (2019) "Fast and wild: Bootstrap inference in Stata using boottest," *Stata Journal*, 19(1): 4-60.

- Rondeau, D. and List, J. (2008), "Matching and Challenge Gifts to Charity: Evidence from Laboratory and Natural Field Experiments," *Experimental Economics*, 11: 253-267.
- Rothman, R., Housam, R., Weiss, H., Davis, D., Gregory, R., Gebretsadik, T., Shintani, A., and Elasy, T. (2006) "Patient understanding of food labels: The role of literacy and numeracy," *American Journal of Preventing Medicine*, 31(5): 391–398.
- Rustichini, A. and Vostroknutov, A. (2008) "Competition with skill and luck," Mimeo, University of Minnesota.
- Rudman, L.A., Greenwald, A.G., and McGhee, D.E. (2001). "Implicit self-concept and evaluative implicit gender stereotypes: Self and ingroup share desirable traits." *Personality and Social Psychology Bulletin* 27(9):1164–1178.
- Ryan K.E. and Ryan A.M. (2005) "Psychological Processes Underlying Stereotype Threat and Standardized Math Test Performance" *Educational Psychologist*, 40: 53-63.
- Shang, J. and Croson, R. (2009), "A Field Experiment in Charitable Contribution: The Impact of Social Information on the Voluntary Provision of Public Goods," *Economic Journal*, 119(540): 1422-1439.
- Sorhaindo, A., and Feinstein, L. (2006) "What is the relationship between child nutrition and school outcomes? Wider benefits of learning," *Research Report No. 18, Institute of Education, University College London*.
- Steele C.M. and Aronson J. (1995) "Stereotype threat and the intellectual test performance of African Americans", *Journal of Personality and Social Psychology*, 89: 797-811.
- Thorndike A., Sonnenberg L., Riis J., Barraclough S., and Levy D. (2012) "A 2-phase labeling and choice architecture intervention to improve healthy food and beverage choices," *American Journal of Public Health*, 102(3): 527–33.
- Thorndike A., Sonnenberg L., Riis J., Barraclough S., and Levy D. (2014) "Traffic-lights labels and choice architecture: Promoting healthy food choices," *American Journal of Preventing Medicine*, 46(2): 143-149.
- Tsukayama, E., Duckwoth, A., and Kim, B. (2013) "Domain-specific impulsivity in school-age children," *Developmental Science*, 16(6): 879-893.
- Usual Dietary Intakes: Food Intakes, U.S. Population, 2007-10. Epidemiology and Genomics Research Program website. National Cancer Institute.

- <http://epi.grants.cancer.gov/diet/usualintakes/pop/2007-10/>. Updated April 24, 2018. Accessed September 6, 2018.
- Volpp, K.G., John, L.K., Troxel, A.B., Norton, L., Fassbender, J., and Loewenstein, G., (2008), “Financial incentive-based approaches for weight loss: A randomized trial,” *Journal of the American Medical Association*, 300 (22), 2631–2637.
- Volpp, K.G., Troxel, A.B., Pauly, M.V., Glick, H.A., Puig, A., Asch, D.A., and Audrian-McGovern, J. (2009) “A randomized controlled trial of financial incentive for smoking cessation,” *New England Journal of Medicine*, 360(7): 699-709.
- Vyth E., Steenhuis I., Heymans M., Roodenburg A., Brug J., and Seidell J. (2011) “Influence of placement of a nutrition logo on cafeteria menu items on lunchtime food choices at Dutch work sites,” *Journal of the American Dietetic Association*, 111: 131–136.
- Weinreb, L., Wehlrer, C., Perloff, J., Scott, R., Hosmer, D., Sagor, L., and Gundersen, C. (2002) “Hunger: its impact on children’s health and mental health,” *Pediatrics*, 110(4): e41.
- Whitaker, R., Phillips, S., and Orzol, S. (2006) “Food insecurity and the risks of depression and anxiety in mothers and behavior problems in their preschool-aged children,” *Pediatrics*, 118(3): e859-e868.
- Winter, J. (2013) “Using the Student’s t-test with extremely small sample sizes,” *Practical Assessment, Research & Evaluation*, 18(10): 1-12.
- Wisdom J, Downs J, and Loewenstein, G. (2010) “Promoting healthy choices: information versus convenience,” *American Economic Journal: Applied Economics*, 2(2): 164–178.
- Wiswall, M., and Zafar, B. (2015) “Determinants of College Major Choice: Identification using an Information Experiment.” *Review of Economic Studies*, 82 (2): 791-824.
- World Health Organization (2009), *Global Health Risks: Mortality and Burden of Disease Attributable to Selected Major Risks*, World Health Organization, Geneva.
- Xie, Y., and Shauman, K. (2003). *Women in Science: Career Processes and Outcomes*. Harvard University Press.
- Zafar, B. (2013) “College Major Choice and the Gender Gap.” *Journal of Human Resources*, 48 (3): 545-595.

5.APPENDIX

APPENDIX CHAPTER 1

Appendix A. Experimental instructions

A1. Instructions for workers. Part one. *Math-Information Treatment*

1. There are two different roles assigned in this experiment. A role can be either worker or firm.
2. Your role is worker.
3. Your task for this part will be the following.
 - a. You belong to a group that is composed of five workers (including yourself) and two firms.
 - b. There will be a market in which the firms have to decide which worker to hire. Each firm will hire only one worker. That means that in your group there will be three workers that will not be hired.
 - c. If the firm hires you, the money you and the firm will make will depend on your performance in the following task. You will be adding a series of five two-digit numbers for five minutes. In this case, the piece rate that you and the firm receive per correct sum is £1.5.
 - d. If the firm does not hire you, your payoff will still depend on your performance. However, this time the piece rate you receive is £0.5 per correct sum. You do not generate anything for the firm.
 - e. The firms do not know anything about you when they are hiring. The only information they will receive will be an avatar that is representing you. The other workers will be also represented by an avatar on the screen.
 - f. Your decision will be to choose how you want to be seen by the firms that are going to hire. So, you will have to decide the avatar that represents you in the market. You will be presented with three options and you have to pick the one you prefer.
4. The firms do NOT know that you yourself chose the avatar that represents you and that the other workers themselves also chose the avatar that represents them.
5. After you and the other workers choose your avatars, the firms will see all the workers that they can hire and will decide who to hire.

6. Once everyone is finished, the experimenter will call you using the number of your computer one by one and pay you privately the amount generated in this experiment.

A2. Instructions for firms. Part one. *Math-Information Treatment*

1. There are two different roles assigned in this experiment. A role can be either worker or firm.
2. Your role is firm.
3. Your task for this part will be the following.
 - a. You belong to a group that is composed of five workers and two firms (including yourself).
 - b. There will be a market in which the firms have to decide which worker to hire. Each firm will hire only one worker. That means that in your group there will be three workers that will not be hired.
 - c. The money that you make in this experiment depends on the performance of the worker that you hire.
4. In order to hire, you will be shown a table with all the workers that have been randomly assigned to your group.
5. Once the hiring process is over, the worker you hired will perform a task consisting in a series of sums (five numbers of two digits) during five minutes. You will earn £1.5 per correct sum the worker gets.
6. Once everyone is finished, the experimenter will call you using the number of your computer one by one and pay you privately the amount generated in this experiment.

A3. Instructions for workers who were hired. Part two. *Math-Information Treatment*

1. Congratulations, you have been hired by the firm.
2. For this part of this experiment you will perform a series of sums (five numbers of two digits) during five minutes.
3. You will be paid a piece rate of £1.5 per correct sum you get. In the same way, the firm also receives £1.5 per correct sum you do.

A4. Instructions for workers who were not hired. Part two. *Math-Information Treatment*

1. You were not hired by the firm.
2. For this part of this experiment you will perform a series of sums (five numbers of two digits) during five minutes.
3. You will be paid a piece rate of £0.5 per correct sum you get. The firm does not make any money from your performance.

A5. Instructions for workers. Part one. *Math-No-Information Treatment (the instructions for firms and for the second stage of the experiment were the same as in MIT)*

1. There are two different roles assigned in this experiment. A role can be either worker or firm.
2. Your role is worker.
3. Your task for this part will be the following.
 - a. You belong to a group that is composed of five workers (including yourself) and two firms.
 - b. There will be a market in which the firms have to decide which worker to hire. Each firm will hire only one worker. That means that in your group there will be three workers that will not be hired.
 - c. The firms do not know anything about you when they are hiring. The only information they will receive will be an avatar that is representing you. The other workers will be also represented by an avatar on the screen.
 - d. Your decision will be to choose how you want to be seen by the firms that are going to hire. So, you will have to decide the avatar that represents you in the market. You will be presented with three options and you have to pick the one you prefer
4. The firms do NOT know that you yourself chose the avatar that represents you and that the other workers themselves also chose the avatar that represents them.
5. After you and the other workers choose your avatars, the firms will see all the workers that they can hire and will decide who to hire.
6. Later, you will participate in a second stage. If you are hired by the firm, you can make three times more money than if you are not hired. We will explain this with more detail later.

7. Once everyone is finished, the experimenter will call you using the number of your computer one by one and pay you privately the amount generated in the experiment.

A6. Instructions for the boxes experiment.

You are now taking part in an economic experiment. Depending on your decisions you will be able to earn money. These instructions describe how you can earn money. Please read them carefully.

The money you will make in this experiment will depend on the performance of somebody else on a different task. The task the other person did is as follows. People were shown 15 photographs depicting individual's faces. For each image, they were asked to identify the emotion depicted on each individual's face. The emotions in the images have been professionally classified by psychologists doing research in this field. The images were projected on the computer screen for a very short period of time (2 seconds). After the image was shown, participants were given four options of emotions from which they selected the option they thought to be correct. They had 20 seconds to submit their answer.

At the front desk of the lab, you will see two boxes. One of the boxes is labeled "males" and the other one is labeled "females". Each box contains 20 slips of paper corresponding to 20 individuals who did the task explained above. Each slip has a number printed on it that corresponds to the number of correct emotions identified by an individual who previously did the task.

Your decision involves choosing the box from which to take out one slip of paper. The number that is printed in your paper will determine your payoff. You will get £0.5 times the number of emotions correctly identified by the person you chose.

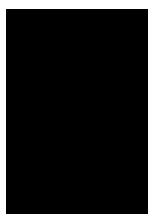
Appendix B. Avatar alternatives



Female Avatar

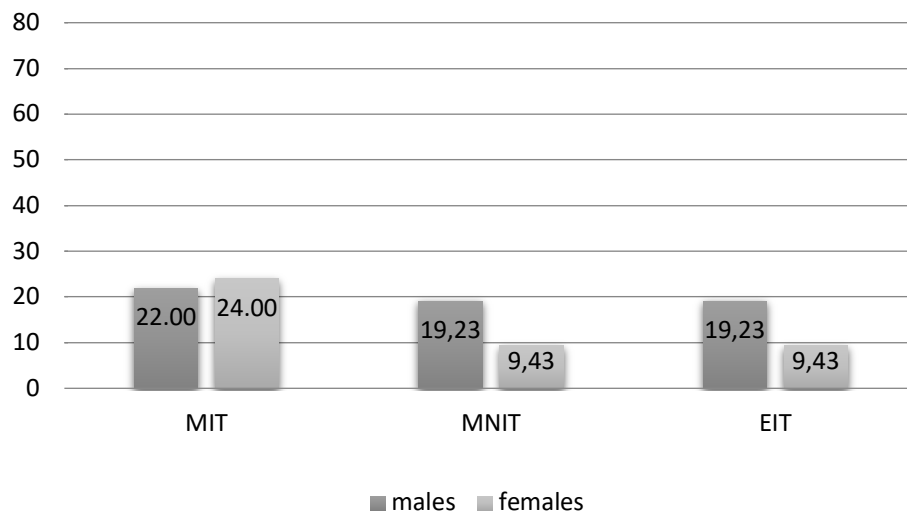


Male Avatar



Neutral Avatar

Appendix C. Neutral Avatar choices



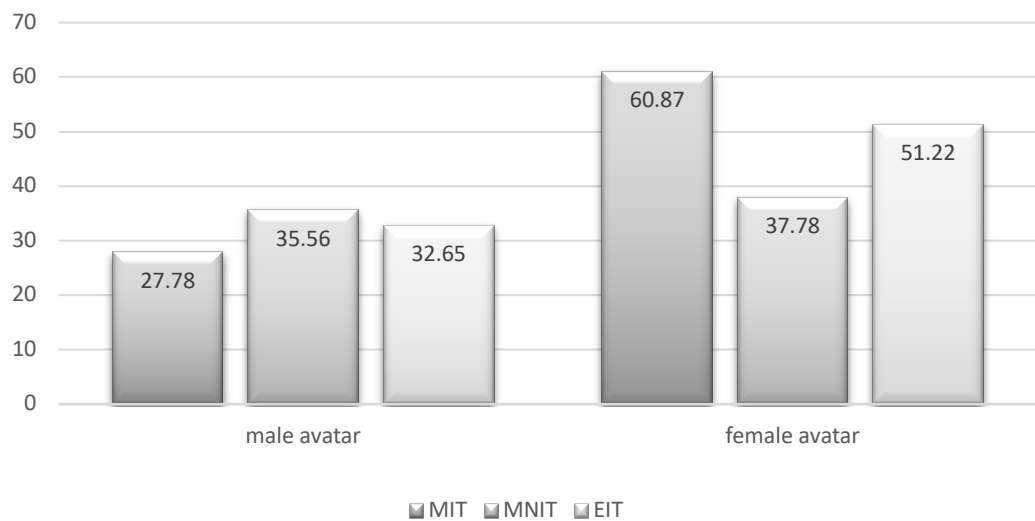
Differences in the percentage of men and women choosing the neutral avatar are not statistically significant in any of the treatments ($Z = -0.236$, $p = 0.813$; $Z = 1.427$, $p = 0.153$; and $Z = 1.427$, $p = 0.153$, two-tailed Mann-Whitney test for *MIT*, *MNIT* and *EIT*, respectively).

Appendix D. Avatar hiring rates

The approach to study hiring decisions provided in the main text does not take into account the supply of avatars that the firms face. Another methodology defines hiring rates in terms of the proportion hired from the available supply of avatars. Figure D1 plots the probability that a worker who chose a particular avatar (male or female) is hired. We do that by computing, from all the male and female avatars that were chosen by the workers, the proportion of them that were hired in the labor market.

As observed in Figure D1, in the *MIT* treatment, the hiring rate for male avatars is 27.78% (there are 54 male avatars, of which 15 are hired); in *MNIT*, this hiring rate is 35.56% (16 of 45). There is no significant difference between these rates ($Z = 0.831$, $p = 0.406$, test of proportions). The hiring rate for female avatars are 60.87% (14 of 23) in *MIT*, and 37.78% (17 of 45) in *MNIT*. The difference between these rates is marginally significant ($Z = 1.809$, $p = 0.071$, test of proportions). Note that firms have the same information in both cases, so that there is an argument for pooling these data in any case. If we do so, the conditional hiring rates are 31.3% (31 of 99) for male avatars and 45.6% (31 of 68) for female avatars; this difference is marginally significant ($Z = 1.876$, $p = 0.061$, test of proportions).

Figure D1: Probability of being hired in each treatment, by avatar



Turning to the emotion-recognition task, the hiring rates are 32.65% (16 out of 49) for male avatars and 51.22% (21 out of 41) for the female avatars. This difference is marginally significant ($Z = 1.783$, $p = 0.075$, test of proportions).

Overall, this analysis does not show any hiring discrimination against female avatars. To the extent that there are differences, we find marginally-significant discrimination *in favor of* female avatars in both the adding-numbers and emotions-recognition tasks. In fact, if we pool the data from all three treatments, we see that 33.78% (50 of 148) of male avatars were hired and 47.71% (52 of 109) female avatars were hired; this difference is statistically significant ($Z = 2.255$, $p = 0.024$). So, female avatars seemed to be more attractive hires than male avatars.

A third form of analysis considers another way that the distribution of avatars can affect hiring patterns. The lower proportion of male avatars hired by firms that we report before in this section could reflect two different issues. While it could be the case that female avatars are *per se* more attractive to firms than male avatars, it could also be that there are more male avatars chosen than those that could be hired (recall that only two of five avatars could be hired), introducing

some sort of bias in our previous results.⁷³ In order to control for this, we calculate, for each combination of avatars, the number of male and female avatars that were hired as a fraction of the maximum number that could be hired.⁷⁴ Figure D2 plots the fraction of avatars (male or female) that are hired, controlling for the particular composition of the group.

Pooling data from *MIT* and *MNIT*, the fraction of male and female avatars hired by companies, as a percentage of the maximum number of avatars that could be hired given the composition of the group, was 41%, and 55%, respectively.⁷⁵ In *EIT*, the respective figures are 38% and 60% for male and female avatars, respectively.⁷⁶ As with the previous analysis, the results indicate that there is positive discrimination towards female avatars.

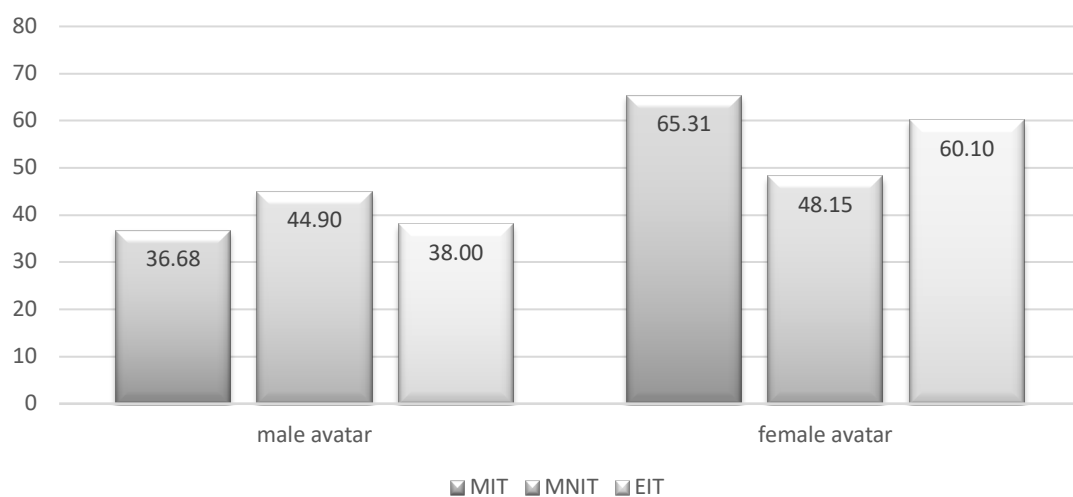
⁷³ In fact, such an overabundance of one avatar type did occur. In *MIT*, there are 10 cases in which there were more than two male avatars available, and two cases with more than two female avatars. In *MNIT*, there are eight cases in which there were more than two male avatars available, and seven cases with more than two female avatars. In *EIT*, there are nine cases in which there were more than two male avatars available, and five cases with more than two female avatars. Note that there are far more cases of more than two male avatars than more than two female avatars.

⁷⁴ For example, consider a group composed of three male avatars, one female avatar and one neutral avatar. The maximum number of male avatars that could be hired in that group would be two (assuming that both firms hire a male avatar), the maximum number of female avatars that could be hired in that group is one, the same as the maximum number of neutral avatars. If in this case, for example, one male avatar and one neutral avatar were hired, we would say that the fraction of male avatars hired in that group is 50% and the fraction of female avatars hired in that group is 100%.

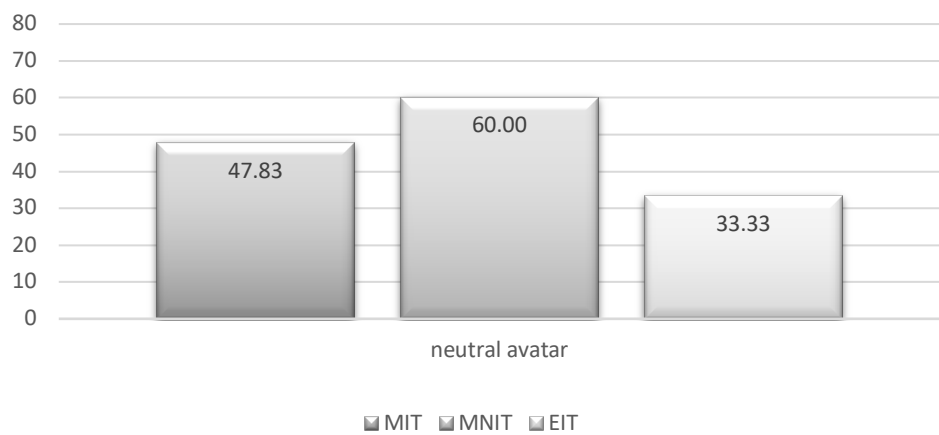
⁷⁵ Differences are statistically significant ($p = 0.048$, paired t -test).

⁷⁶ Differences are statistically significant ($p < 0.001$, paired t -test).

Figure D2: Fraction of avatars hired, by treatment



Appendix E. Neutral avatar hiring rates, by treatment



Wo

Appendix A. Experimental instructions

A1. Anagrams task instructions (group task).

Welcome to our experiment.

The experiment will be divided into two different activities. First, you will participate in the activity that is explained below. Once you finish it, you will receive the instructions for the second activity.

There are 6 participants in the session. These participants have been initially randomly subdivided into two groups of 3 subjects, each determined by the number in the label (odd numbers belong to “Orange Group” and even numbers to “Green Group”).

The first activity involves creating existing words from a number of letters in order to generate money. Each group will be handed a sheet with some letters written on it. The group has 4 minutes to make as many words as possible with at least 3 letters. When the 4 minutes are over the group will receive another sheet with new letters and the former one will be removed. There will be a total of 3 different sheets with letters.

The money you earn will depend on the number of words found by your group. If your group finds a total of **24** words or more, your group will earn **£54**. If the group finds a total number of words smaller than 24, the group will make £1 per word found. The total amount generated will be divided equally among the members of the group.

Proper names are not allowed. Plurals are allowed. Verbal conjugation is allowed. You can use each letter as many times as this letter appears in the word sheet.

A2. Donation stage instructions (group donation).

For the second part of the experiment you will still belong to the same group as in part 1 (either **Orange** or **Green**). In this second part, you will have to decide how much of the money that you obtained in the first part you want to donate to Doctors Without Borders.*

All individual donations within each group will be added and the resulting total amount of each group will be transferred to Doctors Without Borders.

You already received the money you made in the first part of the experiment. You have also received an empty envelope. Your task is to introduce in the envelope all the money you want to donate to Doctors Without Borders. Once you have introduced the money into the envelope, you just have to put the envelope inside the box that is located in the table close to the door and you can leave the experiment with the cash you kept for yourself.

To show that the money donated in the session has been actually transferred to the charity, we can send you an email with the receipt from Doctors Without Borders once the donation has been made. Please, provide your university email address as clear as possible to the experimenter if you want to receive proof of the transfer.

☐

I want to receive proof of the transfer. My university email address is

*Doctor Without Borders is a medical and humanitarian international organization that provides help to the victims of natural or human disasters and armed conflicts without any kind of race, sex, religion, philosophy or political discrimination.

Appendix A: Justification of trays

The diet we follow is one of the most important factors in determining our health. Because of this, it is important to follow a healthy diet, in accordance with the needs of each life stage.

In this study, four trays of prepared items which had not undergone any culinary treatment were used. The composition of each tray was carried out with items in accordance with the targeted population.

Trays with 0 points and 2.5 points:

The content of these two trays is ultra-processed products which each share low-quality ingredients. Among these ingredients, we find refined flours, non-virgin vegetable oils, added sugar and salt. All of these are associated with illnesses such as high blood pressure, diabetes, depression, obesity and cancer, among others.

Refined flours are those in which the refining process has removed the bran from them and the germ from the whole grain. In this way, fibres, vitamins, minerals, anti-oxidants and phytosterols are lost; and, as such, the nutritional value is lowered. This occurs with sliced bread which, in addition, has sugar added to it.

The same refining process occurs with vegetable oils. In particular, in the products that are part of these trays, we find palm oil. This oil is related to an increase in cardiovascular risk, and visceral fat. It is also related to certain types of cancer.

Considering that the WHO (World Health Organization) recommends that the ingestion of sugar in children does not exceed 15 grams per day, we can see how the consumption of one Phoskitos or two Bollycaos would exceed these

recommendations, risking the health of a school-going child. The same occurs with sweets, Nocilla sticks, cereals, milkshakes and similar products.

In the case of cold meats, although their refined sugar and salt content is lower, they still present a high content of salt and bad quality fats. Processed meats are related to some types of cancer and other pathologies.

As such, all of these products have a heightened calorie density and very few nutrients. Therefore, they only provide empty calories, what we refer to as “low nutritional density”.

The items present in tray 0 do not contain healthy ingredients, and they have empty calories. These products belong to the lowest tray because not only are the responses they cause in the organism non-beneficial but their consumption may result in serious health problems.

In tray 2.5, we find items that are not recommended either. However, in certain circumstances, they may be used as they contain some raw materials that do provide valuable nutrients such as protein in dairy products, minerals and vitamins in cereals and sources of protein and iron, such as chorizo or salchichon.

Tray with 5 points:

This tray is found in the middle of the trays in this study, since it has properties of the two previous trays and the two last ones.

This tray contains items such as milk which provides nutrients such as beneficial fatty acids, vitamins such as A, D, C and those from the B group, in addition to minerals such as calcium, phosphorus and potassium.

Bread in this tray provides fiber by being whole and less refined.

With respect to the yoghurts of this tray, note that they are high in previously-mentioned sugars, this being the reason for not including them in subsequent trays.

Finally, “Babybell” cheese and the turkey used to prepare sandwiches, contain different added substances which, over a longer period of time, may have repercussions for our health.

Summarizing, in the tray with 5 points we have products that are preparations with more nutritional value but the advantages of their consumption still could be improved by a wide range of other items.

Tray 7.5 and 10 points:

These trays are formed by minimally-processed dairy products, fruits, dried fruits, whole rye bread, jamón serrano and virgin olive oil.

They are characterized by a high content of total fiber, which relieves hunger, regulates intestinal motility and bacterial micro-flora substrates of the colon. They prevent illnesses such as constipation, diverticulosis, inflammatory bowel disease and irritable bowel syndrome, cancer, diabetes mellitus, atherosclerosis and obesity, among others.

Dried fruit has a good protein content. They also have a high fat content which provides energetic and healthy properties due to the mono-saturated and poly-saturated fatty acids that they are made of. Nuts and almonds are the best choices.

Fruits are a real dietary jewel. They are characterized by having a high content of simple sugars but, unlike ultra-processed items, these, by being found inside the natural source, do not cause any of the previously-mentioned health problems. Fruits used in these trays do not contain fats. The vitamins that are present in fruits and vegetables, such as Vitamin C, make them very important.

We include hard foodstuffs such as apples, that are appropriate for reinforcing teeth, providing Magnesium, Calcium and Fluoride in varied and

balanced diets. In addition, water is the essential drink, providing schoolchildren with Fluoride in addition to other minerals.

The dairy products of this group are strongly recommended as they help to battle different infections that may arise, helping to regenerate intestinal flora and to maintain a good system of defenses.

The last tray has a score of 10 points as we find ingredients of high nutritional value, such as olive oil, the main ingredient of our Mediterranean diet. It is rich in vitamins A, D, E and K. Olive oil provides us with a wide range of benefits of which the following stand out: protection from vascular illnesses, expels grassy residues from the liver, anti-oxidant action, improvement of the digestive system, reducing, additionally, constipation, improvement of metabolic functions and cerebral development, stimulating the absorption of certain minerals such as calcium and controlling the level of glucose in blood among others.

The sandwich included in this tray is more complete than in previous baskets and with better ingredients: whole rye bread, jamón serrano (and not processed meat such as chorizo or turkey), and virgin olive oil.

Appendix B. Summary of nutritionist's talk

This talk aimed at schoolchildren has as their main objective to help the child to differentiate real food and ultra-processed foods (pastries, snacks...). Once this difference has been established, we relate the state of health with the consumption of one or the other food, focusing on the following aspects of the ultra-processed ones (the information is transmitted using an easy-to-understand language adapted to their ages):

- Ultra-processed products have a high caloric density and produce an increase in weight.
- Ultra-processed products generate a feeling of hunger shortly after their intake, so the tendency will be to eat again.
- Ultra-processed products produce pleasurable stimuli at the neurological level, so our brain will always prefer them over real food.
- Ultra-processed products have very intense and enhanced flavors compared to real food, so they will be the main choice for eating, leaving healthy foods aside.

Once the problem is presented, the benefits of real food in health are listed. In particular, we focus on the advantages of a high consumption of fruits and vegetables:

- In general, they produce or maintain an adequate state of health.
- They generate satiety, which prevents us from eating between hours.
- They provide essential nutrients such as vitamins and minerals.
- The intake of vegetables in the diet provides a great source of fiber, which allows good intestinal health.

Appendix C. Composition of food trays



Appendix D.1. Socio-economic survey

1. This questionnaire has been filled by:

- a. Father
- b. Mother
- c. Both
- d. Other

2. The father has a University degree:

- a. Yes
- b. No

3. The mother has a University degree:

- a. Yes
- b. No

4. How many of the following items does the family own?

Please, indicate for each item

	0	1	2	3	4	5	6	7 or more
Computers								
Laptops								
Tablets								
Video games								
Electronic books								
Cars or motorbikes								
Bathrooms in the house you are current living								
Number of properties (on top of the one you are currently living)								

5. Please, indicate your level of participation in your child's activities

Please, indicate for each activity

	Never	Some days	Often	Every day
We encourage him/her to study				
We ask about his/her homework				

We check he/she does his/her homework				
We ask about his/her day in school				
We help with his/her homework				

Appendix D.2. Impulsivity questionnaire

Interpersonal Impulsivity. How often...

...do I interrupt other people?

...do I say something rude?

...do I lose temper?

...do I talk back when upset?

1 = almost never, 2 = about once per month, 3 = about 2 to 3 times per month, 4 = about once per week, and 5 = at least once per day

Schoolwork Impulsivity. How often...

...do I forget something needed for school?

...do I cannot find something because of mess?

...do I not remember what someone said to do?

1 = almost never, 2 = about once per month, 3 = about 2 to 3 times per month, 4 = about once per week, and 5 = at least once per day

Appendix D.3. Eating habits survey

WEEKLY CONSUMPTION					
	NEVER	ONCE A WEEK	2/3 TIMES A WEEK	ALMOST EVERYDAY	ON A DAILY BASIS
	DAIRY				
Milk					
Cheese and butter					
Desserts: yogurt, custard, flan, pudding, rice pudding...					
Ice cream					
	EGGS, MEAT, FISH				
Eggs					
Chicken, turkey, rabbit					
Red meat: beef, pork					
Serrano ham, ham, salchichon, chorizo, sausage...					
White fish					
Oily fish					
Canned fish					
	VEGETABLES				
Vegetables					
	FRUITS				
Fruits					
Olives and avocados					
	DRIED FRUITS				
Almonds, peanuts, pistachios, nuts...					

	DRIED VEGETABLES				
Lentils, chickpeas...					
	RICE AND CEREALS				
White bread					
Wholemeal bread					
Breakfast cereals					
White rice					
Pasta					
Wheat pasta					
	OILS				
Extra virgin olive oil					
Other types of olive oil					
Sunflower oil					
Margarine					
	PASTRY				
Supermarket pastry					
Homemade pastry					
Chocolates					
	DRINKS				
Water					
Soft drinks					
Fresh squeezed orange juice					
Bottle juices					
	OTHERS				
Fast food					

Ketchup, mayo					
Honey, sugar					

Appendix E. Additional Results

**Table E.1 – Descriptive statistics of pre-treatment characteristics
across treatment groups**

	Baseline	NT	GT	PT
Dietary habits (weekly consumption)^a				
Milk and dairy products	4.129	4.028	3.972	3.982
Eggs	2.516	2.714	2.805	2.744
Meat (beef, pork and chicken)	3.145	3.073	2.944	2.976
Seafood	2.139	2.222	1.982	2.079
Fruit and vegetables	2.910	2.980	2.794	2.914
Bread	2.913	2.947	2.914	2.987
Rice and Cereals	2.428	2.310	2.186	2.140
Cooking Oils	3.001	3.023	3.075	2.990
Confectionery	1.912	1.952	1.914	1.913
Other fats and sugars	2.140	2.012	2.097	1.942
Demographic and socio-economic characteristics				
Parents hold University degree ^b	0.032	0.046	0.056	0.080
Both parents are economically inactive ^c	0.041	0.031	0.028	0.022
Household Income ^d	26.221	35.044	28.692	30.611
Parents' involvement in children's school success ^e	3.5375	3.5621	3.4736	3.4583
Average school grade ^f	7.241	7.097	7.214	7.843
Frequency of child's school absenteeism ^g	0.053	0.044	0.093	0.020
Proportion of male subjects ^c	48.681	56.338	53.488	52.702

Notes:

^a Average number of times that the subjects allocated to each treatment eat items in the different food categories in a typical week.

^b Proportion of subjects allocated to each treatment for whom at least one of their parents holds a university degree.

^c As a proportion of all subjects allocated to each treatment.

^d Proxied by the percentage of households that own the following goods: house, car, desktop computer, notebook computer, and tablet.

^e Index based on parents' self-reported attention to children's class attendance and homework and parent's overall involvement in their children's scholastic work.

^f Coded on a scale from 1 to 10.

^g Coded as 1 if frequent, 0 otherwise. Based on the teachers' judgment.

**Table E.2 - Tests for differences in pre-treatment characteristics
across treatment groups**

	GT - Baseline	NT – Baseline	PT – Baseline	GT-NT	GT-PT	NT-PT
Dietary habits (weekly consumption)						
Milk and dairy products	Z = -0.748 p = 0.648	Z = -0.457 p = 0.455	Z = -1.579 p = 0.111	Z = -0.201 p = 0.840	Z = -0.726 p = 0.468	Z = 0.851 p = 0.395
Eggs	Z = 1.434 p = 0.152	Z = 0.852 p = 0.394	Z = 1.263 p = 0.207	Z = 0.586 p = 0.558	Z = 0.471 p = 0.637	Z = -0.220 p = 0.8260
Meat (beef, pork and chicken)	Z = -1.045 p = 0.296	Z = -1.262 p = 0.207	Z = -0.939 p = 0.348	Z = 0.437 p = 0.662	Z = -0.215 p = 0.830	Z = 0.660 p = 0.509
Seafood	Z = -1.031 p = 0.302	Z = 0.496 p = 0.620	Z = 0.479 p = 0.632	Z = -1.570 p = 0.116	Z = -0.899 p = 0.369	Z = 1.339 p = 0.180
Fruit and vegetables	Z = -0.838 p = 0.402	Z = 0.028 p = 0.978	Z = -0.556 p = 0.578	Z = -0.717 p = 0.473	Z = -1.441 p = 0.150	Z = 0.438 p = 0.669
Bread	Z = -0.01 p = 0.998	Z = -0.134 p = 0.894	Z = -0.366 p = 0.714	Z = -0.155 p = 0.877	Z = 0.419 p = 0.675	Z = -0.229 p = 0.818
Rice and Cereals	Z = 0.011 p = 0.998	Z = 0.061 p = 0.951	Z = 0.701 p = 0.484	Z = -0.512 p = 0.588	Z = 0.367 p = 0.713	Z = 0.808 p = 0.419
Cooking Oils	Z = -0.405 p = 0.685	Z = -0.116 p = 0.907	Z = 0.075 p = 0.940	Z = 0.405 p = 0.685	Z = 0.832 p = 0.405	Z = 0.399 p = 0.690
Confectionery	Z = -0.985 p = 0.325	Z = -0.641 p = 0.522	Z = 1.570 p = 0.116	Z = -0.227 p = 0.820	Z = 0.782 p = 0.434	Z = 0.875 p = 0.381
Other fats and sugars	Z = -0.189 p = 0.850	Z = -1.590 p = 0.112	Z = -1.534 p = 0.125	Z = -1.581 p = 0.114	Z = 1.453 p = 0.146	Z = 0.564 p = 0.573
Demographic and socio-economic characteristics						
Parents hold University degree	Z = 0.401, p = 0.688	Z = 1.110, p = 0.266	Z = 0.957, p = 0.339	Z = 1.340, p = 0.181	Z = -1.758, p = 0.078	Z = -0.642, p = 0.532
Both parents are economically inactive	Z = -1.017, p = 0.309	Z = -1.226, p = 0.220	Z = -2.000, p = 0.045	Z = -0.061, p = 0.951	Z = 1.171 p = 0.242	Z = 1.222, p = 0.222
Household Income	Z = 0.450, p = 0.653	Z = 1.803, p = 0.070	Z = 0.693, p = 0.488	Z = -1.100, p = 0.271	Z = -0.288, p = 0.773	Z = 0.637, p = 0.524
Involvement in children's school success	Z = -0.778, p = 0.436	Z = 0.380, p = 0.704	Z = -1.007, p = 0.313	Z = -1.165, p = 0.244	Z = 0.167, p = 0.867	Z = 1.420, p = 0.156
Average school grade	Z = -0.099, p = 0.921	Z = -0.497, p = 0.619	Z = 1.461, p = 0.144	Z = 0.350, p = 0.726	Z = -1.609, p = 0.110	Z = -1.967, p = 0.049
Frequency of child's school absenteeism	Z = 0.841, p = 0.400	Z = -1.299, p = 0.194	Z = -1.990, p = 0.146	Z = 2.004, p = 0.125	Z = 2.658, p = 0.078	Z = -1.010, p = 0.311
Proportion of male subjects	Z = 0.501, p = 0.616	Z = 0.925 p = 0.355	Z = 0.490, p = 0.624	Z = -0.295, p = 0.768	Z = 0.082, p = 0.935	Z = 0.437, p = 0.661

Joint test for all covariates	$\chi^2= 5.230$, p=0.514	$\chi^2= 10.700$, p=0.151	$\chi^2= 6.670$, p=0.464	$\chi^2= 10.101$, p=0.185	$\chi^2= 4.581$, p=0.710	$\chi^2= 5.941$, p=0.546

Note: Covariate balance checks based on the test statistics proposed by Hansen and Bowers (2008) to assess difference in individual covariates and combined differences in all characteristics (bottom row of the table) across treatment groups, under cluster-level randomization. The coding of the pre-treatment characteristics is the same as in Table E.1. The p-values correspond to two-sided tests.

Table E.3 – Multinomial logit model for treatment assignment

Treatment	Covariate	Estimates
<i>Grades (GT)</i>	<i>Constant</i>	-2.97 (1.79)
	Cluster-Adjusted p-values	(0.50)
	<i>Milk and dairy products</i>	-5.20 (3.29)
	Cluster-Adjusted p-values	(0.51)
	<i>Carbohydrates</i>	-1.95 (1.93)
	Cluster-Adjusted p-values	(0.49)
	<i>Proteins</i>	2.75 (2.90)
	Cluster-Adjusted p-values	(0.11)
	<i>Fruit and vegetables</i>	-5.41 (4.09)
	Cluster-Adjusted p-values	(0.14)
	<i>Fats and sugars</i>	1.40 (1.35)
	Cluster-Adjusted p-values	(0.16)
	<i>Male</i>	-0.42 (0.50)
	Cluster-Adjusted p-values	(0.44)
	<i>Parents hold University degree</i>	0.24 (0.76)
	Cluster-Adjusted p-values	(0.60)
	<i>Both parents are economically inactive</i>	-0.77 (1.26)

	Cluster-Adjusted p-values	(0.67)
	<i>Household Income</i>	0.32 (0.50)
	Cluster-Adjusted p-values	(0.58)
	<i>Involvement in children's school success</i>	0.53 (0.38)
	Cluster-Adjusted p-values	(0.50)
	<i>Frequency of child's school absenteeism</i>	1.90 (1.25)
	Cluster-Adjusted p-values	(0.50)
	<i>Average school grade</i>	0.17 (0.11)
	Cluster-Adjusted p-values	(0.71)
<i>Nutritionist (NT)</i>	<i>Constant</i>	-1.41 (1.44)
	Cluster-Adjusted p-values	(0.72)
	<i>Milk and dairy products</i>	-5.15 (3.10)
	Cluster-Adjusted p-values	(0.51)
	<i>Carbohydrates</i>	-2.08 (3.10)
	Cluster-Adjusted p-values	(0.49)
	<i>Proteins</i>	3.32 (3.50)
	Cluster-Adjusted p-values	(0.11)
	<i>Fruit and vegetables</i>	-4.74 (4.96)
	Cluster-Adjusted p-values	(0.14)
	<i>Fats and sugars</i>	1.37 (2.39)
	Cluster-Adjusted p-values	(0.16)
	<i>Male</i>	-0.25 (0.42)
	Cluster-Adjusted p-values	(0.44)
	<i>Parents hold University degree</i>	0.59 (0.67)
	Cluster-Adjusted p-values	(0.60)
	<i>Both parents are economically inactive</i>	-1.34 (1.04)

	Cluster-Adjusted p-values	(0.67)
	<i>Household Income</i>	1.02 (0.42)
	Cluster-Adjusted p-values	(0.58)
	<i>Involvement in children's school success</i>	0.48 (0.33)
	Cluster-Adjusted p-values	(0.50)
	<i>Frequency of child's school absenteeism</i>	-0.34 (1.49)
	Cluster-Adjusted p-values	(0.50)
	<i>Average school grade</i>	-0.04 (0.14)
	Cluster-Adjusted p-values	(0.71)
<i>Parents (PT)</i>	<i>Constant</i>	-3.89 (1.95)
	Cluster-Adjusted p-values	(0.50)
	<i>Milk and dairy products</i>	-5.54 (3.31)
	Cluster-Adjusted p-values	(0.51)
	<i>Carbohydrates</i>	-2.12 (1.98)
	Cluster-Adjusted p-values	(0.49)
	<i>Proteins</i>	4.57 (2.90)
	Cluster-Adjusted p-values	(0.11)
	<i>Fruit and vegetables</i>	-4.91 (4.08)
	Cluster-Adjusted p-values	(0.14)
	<i>Fats and sugars</i>	1.16 (1.42)
	Cluster-Adjusted p-values	(0.16)
	<i>Male</i>	-0.21 (0.53)
	Cluster-Adjusted p-values	(0.44)
	<i>Parents hold University degree</i>	0.50 (0.78)
	Cluster-Adjusted p-values	(0.60)
	<i>Both parents are economically inactive</i>	0.45 (1.33)

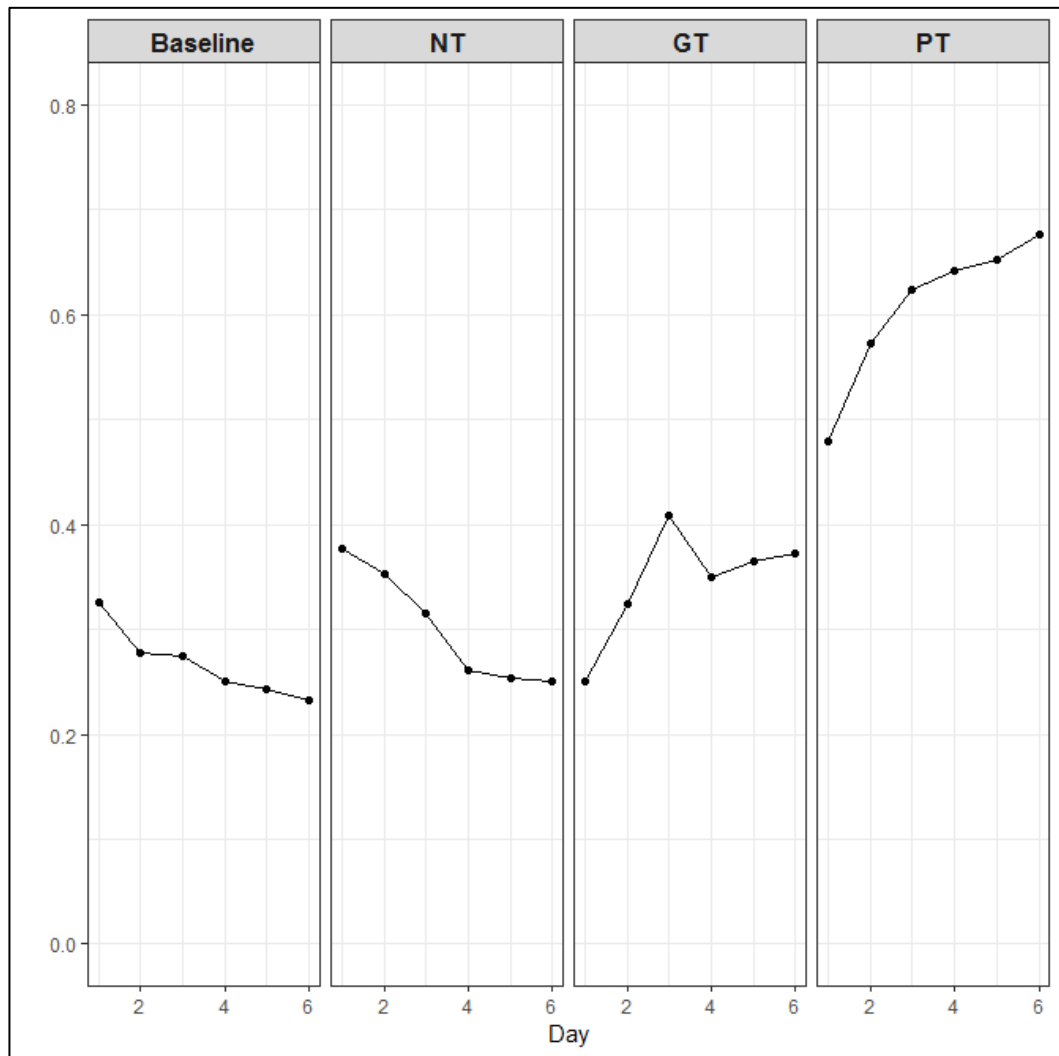
Cluster-Adjusted p-values	(0.67)
<i>Household Income</i>	0.32 (0.55)
Cluster-Adjusted p-values	(0.58)
<i>Involvement in children's school success</i>	0.62 (0.41)
Cluster-Adjusted p-values	(0.50)
<i>Frequency of child's school absenteeism</i>	0.34 (1.48)
Cluster-Adjusted p-values	(0.50)
<i>Average school grade</i>	0.22 (0.20)
Cluster-Adjusted p-values	(0.71)
Likelihood-ratio test	$\chi^2 = 25.61, p = 0.22$

Notes: The table reports parameter estimates from a multinomial logit model for subjects' treatment assignment, with *Baseline* as the reference category. Standard errors are reported in parentheses (first line below the coefficients). To account for intra-school dependence among subjects while taking into consideration the small number of schools in our sample, we used Ibragimov and Muller's (2010) approach to estimate cluster-adjusted t-statistics (see also Esarey and Menger, 2019); p-values for each pre-treatment variable are reported in the table (second line below the coefficients). *Male*, *Parents hold University degree*, *Both parents are economically inactive* and *Frequency of child's school absenteeism* are indicator variables; *Milk and dairy products*, *Carbohydrates*, *Protein*, *Fruit and vegetables* and *Fats and sugars* are the average number of times subjects consume food items in those categories during a typical week; *Household Income*, *Involvement in children's school success* and *Average school grade* are coded as in Table E.2. As seen in the table, none of the covariates has a statistically significant influence on treatment adjustment, once we account for the small number of clusters in our sample. Similarly, the Likelihood ratio test indicates that including these covariates does not improve fit vis-à-vis a null (intercept only) model.

Table E.4 – Proportion of healthy choices and differences across treatments using alternative definitions of “healthy choice”

Treatment	“Healthy choice” = grades equal to 10		“Healthy choice” = grades equal to or greater than 5	
	Proportion of healthy choices	Mann-Whitney tests vis- à-vis <i>Baseline</i>	Proportion of healthy choices	Mann-Whitney tests vis- à-vis <i>Baseline</i>
<i>Baseline</i>	0.27	-	0.49	-
<i>NT</i>	0.30	$p = 0.200$	0.58	$p = 0.050$
<i>GT</i>	0.34	$p = 0.050$	0.59	$p = 0.050$
<i>PT</i>	0.60	$p = 0.050$	0.85	$p = 0.050$

Figure E.1. Robustness: Percentage of healthy choices over time, using a stricter definition of “healthy choice”



Notes: The figure replicates the analysis summarized in Figure 2 of the main text, but defining as “healthy choices” those with grades equal to 10 only.

Table E.5. (Multi-level) regression analysis of subjects' average grades

	Data from intervention period			Four months after intervention
	(1)	(2)	(3)	(4)
<i>Constant</i>	4.65*** (0.33)	5.86*** (0.56)	4.92*** (0.59)	5.02*** (0.38)
<i>GT</i>	1.04** (0.43)	1.22** (0.58)	1.59** (0.58)	0.90* (0.48)
<i>NT</i>	0.82** (0.41)	1.02** (0.47)	0.94* (0.52)	0.59 (0.49)
<i>PT</i>	3.20*** (0.40)	3.14*** (0.41)	3.12*** (0.51)	2.38*** (0.52)
<i>Male</i>		0.28 (0.22)	0.47 (0.26)	
<i>Average school grade</i>		0.16*** (0.06)	0.12 (0.08)	
<i>Personal Impulsivity</i>		-0.64 (0.47)	-1.13** (0.56)	
<i>School Impulsivity</i>		-0.29* (0.15)	-0.22 (0.17)	
<i>Parents hold University degree</i>			-0.14 (0.23)	
<i>Household Income</i>			0.49*** (0.14)	
# Observations	1,378	1,351	945	213
Deviance Information Criterion	5,129.64	5,025.14	3,324.99	146.70

Notes: The table reports parameter estimates from multi-level regression models including subject, school and – in columns (1)-(3) – period (day) random effects. Bayesian inference via Markov chain Monte Carlo simulations is conducted to account for the small number of schools in our sample (Gelman, 2006); the table reports posterior means and standard errors (in parentheses). Significance levels: ***, **, and * denote significance at $p = 0.01$, 0.05 , and 0.10 , respectively.

Table E.6. Probit regression on the probability of choosing healthy food items

	Data from intervention period	Four months after intervention
--	-------------------------------	--------------------------------

	(1)	(2)	(3)	(4)
<i>Constant</i>	-0.12*** (0.02)	0.60*** (0.22)	0.60*** (0.22)	-1.15*** (0.01)
Wild bootstrap-t p-value	(0.00)	(0.00)	(0.00)	(0.00)
<i>GT</i>	0.61* (0.17)	0.70** (0.12)	0.95** (0.17)	0.96* (0.09)
Wild bootstrap-t p-value	(0.06)	(0.03)	(0.03)	(0.06)
<i>NT</i>	0.45* (0.15)	0.53* (0.19)	0.60* (0.26)	0.49*** (0.08)
Wild bootstrap-t p-value	(0.06)	(0.07)	(0.07)	(0.00)
<i>PT</i>	1.94** (0.28)	1.93* (0.28)	2.01*** (0.41)	1.86*** (0.07)
Wild bootstrap-t p-value	(0.02)	(0.06)	(0.00)	(0.00)
<i>Male</i>		0.09 (0.12)	0.11 (0.18)	
Wild bootstrap-t p-value		(0.55)	(0.48)	
<i>Average school grade</i>		0.06 (0.04)	0.02 (0.07)	
Wild bootstrap-t p-value		(0.25)	(0.86)	
<i>Personal Impulsivity</i>		-0.33 (0.23)	-0.45 (0.28)	
Wild bootstrap-t p-value		(0.17)	(0.17)	
<i>School Impulsivity</i>		-0.21* (0.12)	-0.14 (0.12)	
Wild bootstrap-t p-value		(0.09)	(0.24)	
<i>Parents hold University degree</i>			-0.24 (0.17)	
Wild bootstrap-t p-value			(0.30)	
<i>Consumption</i>			0.30 (0.19)	
Wild bootstrap-t p-value			(0.58)	
# Observations	1,378	1,351	945	213
Pseudo-R ²	0.18	0.19	0.23	0.15

Notes: Estimation method is maximum likelihood. Cluster-robust standard errors clustered by school in parentheses (first line below the coefficients). The p-values for two-sided Wald tests computed according to Kline and Santos' (2012) score bootstrap accounting for small number of clusters (schools) are also reported in parentheses (second line below the coefficients). Significance levels (based on the – more conservative, bootstrapped – Wild tests): ***, **, and * denote significance at $p = 0.01$, 0.05 , and 0.10 , respectively.

**Table E.7. Regression analysis of the daily differences in
subjects' grades across treatments**

Covariates	Parameter estimates	Wild bootstrap
Constant	4.72 (0.05)	
<i>PT x Period1</i>	2.84 (0.26)	t=10.9, p=0.01
<i>PT x Period2</i>	3.11 (0.12)	t=24.99, p=0.00
<i>PT x Period3</i>	3.43 (0.21)	t=16.24, p=0.00
<i>PT x Period4</i>	3.37 (0.16)	t=20.78, p=0.00
<i>PT x Period5</i>	3.30 (0.30)	t=11.05, p=0.00
<i>PT x Period6</i>	2.85 (0.60)	t=4.71, p=0.00
<i>GT x Period1</i>	0.39 (0.08)	t=4.98, p=0.16
<i>GT x Period2</i>	1.07 (0.49)	t=2.20, p=0.00
<i>GT x Period3</i>	1.73 (0.34)	t=5.11, p=0.00
<i>GT x Period4</i>	1.11 (0.29)	t=3.85, p=0.05
<i>GT x Period5</i>	1.25 (0.35)	t=3.61, p=0.02
<i>GT x Period6</i>	0.73 (0.40)	t=1.83, p=0.30
<i>NT x Period1</i>	1.65 (0.27)	t=6.11, p=0.03
<i>NT x Period2</i>	0.81 (0.53)	t=1.51, p=0.30
<i>NT x Period3</i>	0.82 (0.21)	t=4.54, p=0.13
<i>NT x Period4</i>	0.52 (0.16)	t=4.54, p=0.15
<i>NT x Period5</i>	0.39 (0.10)	t=2.27, p=0.39
<i>NT x Period6</i>	0.24 (0.17)	t=1.35, P=0.61

Notes: The table reports the parameter estimates from the specification used as a basis to produce Figure 4 in the main text. Column (1) reports the OLS estimates; cluster-robust standard errors clustered by school in parentheses. Column (2) reports t-statistic and p-values (in parentheses) for two-sided Wald tests computed according to Cameron *et al.*'s (2008) wild bootstrap-t procedure accounting for small number of clusters (schools).