Discussion Papers of the
Max Planck Institute for
Research on Collective Goods
2021/1

Improving Healthy Eating in Children: Experimental Evidence

Gary Charness
Ramón Cobo-Reyes
Erik Eyster
Gabriel Katz
Ángela Sánchez
Matthias Sutter

MAX PLANCK
SOCIETY

Discussion Papers of the
Max Planck Institute
for Research on Collective Goods 2021/1

Improving Healthy Eating in Children: Experimental Evidence

Gary Charness / Ramón Cobo-Reyes / Erik Eyster / Gabriel Katz / Ángela Sánchez / Matthias Sutter

January 2021

Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Str. 10, D-53113 Bonn
https://www.coll.mpg.de

Improving healthy eating in children: Experimental evidence*

GARY CHARNESS,[1] RAMÓN COBO-REYES,[2] ERIK EYSTER,[1] GABRIEL KATZ,[3] ÁNGELA SÁNCHEZ,[4] AND MATTHIAS SUTTER[5]

[1] UNIVERSITY OF CALIFORNIA, SANTA BARBARA

[2] AMERICAN UNIVERSITY OF SHARJAH

[3] UNIVERSITY OF EXETER AND UNIVERSIDAD CATOLICA DEL URUGUAY

[4] NYU ABU DHABI

[5] MAX PLANCK INSTITUTE BONN, UNIVERSITY OF COLOGNE, UNIVERSITY OF INNSBRUCK, AND IZA BONN

DECEMBER 27, 2020

Abstract

We present a field experiment to study the effects of non-monetary incentives on healthy food choices of 282 children in elementary schools. Previous interventions have typically paid participants for healthy eating, but this often may not be feasible. We introduce a system where 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.

JEL-Codes: C93, I12

Keywords: Healthy eating, children, parents, non-monetary incentives, field experiment

* We received helpful comments from Kelly Bedard, Michelle Belot, Francisco Lagos, Sofia Monteiro, Heather Royer, Anya Samek, and seminar participants at the American University of Sharjah and UCSB. Assistance provided by David Fuentes, Alberto Ramos, Nuria Losada and Maria Eugenia Mata is greatly appreciated. Financial support from the Max Planck Society, the American University of Sharjah and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany´s Excellence Strategy – EXC 2126/1– 390838866 is gratefully acknowledged. This study is registered in the AEA RCT Registry and the unique identifying number is: AEARCTR-0003370.

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 both their physical well-being and cognitive development. It has been found to weaken the immune system (Sorahindo and Feinstein, 2006) and to contribute to the development of dental caries and diabetes (Noble and Kanoski, 2016). Moreover, poor diet hinders growth and even undermines 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 American children did not eat enough fruit to meet daily recommendations in the period 2007-2010, and 93% did not eat enough vegetables.[1] On a different dimension, the National Health and Nutrition Examination Survey reported in 2011-2012 that over 30% of U.S. children between 2-19 years old were overweight. This percentage has tripled since the survey carried out in 1971-1974. Because dietary habits are typically set during childhood, it is crucial that improvements are made during this period (Haire-Joshu and Tabak, 2016).

Designing, implementing, and evaluating interventions aiming to produce durable changes in children's eating behavior 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. While it has been shown that paying people leads to better performance (Belot *et al,.* 2016; Loewenstein *et al*., 2016) this is not always practical or feasible. In particular, monetary incentives cannot easily and permanently be rolled

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

out on a large scale. Therefore, one should consider what can be viewed as a next generation of behavioral interventions, where no direct financial reward is paid for achieving goals. This is what we do in this paper. Our approach can be implemented with relatively minimal funding and so could be widely adopted.

In this regard, we conduct an innovative field experiment in Spanish elementary schools to study whether and how it might be possible to influence or incentivize a total of 282 school children, aged nine to ten, to make healthier food choices. Our design includes four treatments. In all treatments, children were presented with five different food trays containing an assortment of snacks and made four choices from all snacks across the trays. The intervention lasted for three weeks, during which children chose snacks on two prescheduled days per week. 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* treatment, a clearly-labeled "grade" was associated with each of the trays, with healthier foods receiving higher grades. Finally, in the *Parents* treatment, the children saw the same grades as in the *Grades* treatment, and parents also received weekly reports on their children's average food grade that week; the children knew that their parents would receive this information. Four months after the end of the intervention, 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 healthier choices than the *Baseline*. We also had expected the information released by the nutritionist to significantly change the children's behavior, and so we predicted differences between the *Nutritionist* treatment and the *Baseline*.

We find a modest improvement in healthy eating in both the *Grades* and *Nutritionist* treatments, accompanied by different trends over time. Whereas the proportion of healthy choices is 36% in the *Baseline*, it is 45% in the *Nutritionis*t treatment, and 46% in the *Grades* treatment. The *Parents* treatment has far larger effects, with 74% of the choices being healthy foods—more than double the rate of the *Baseline*. This dramatic increase cannot be attributed 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%, 47%, 54%, and 69% in *Baseline*, *Nutritionist*, *Grades* and *Parents*, respectively. This suggests benefits to these interventions even after removal.

The contribution of this paper is threefold. First, it 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 their children's average grade but not their exact choices. So, while parents certainly influence their children's decisions, they cannot dictate or even precisely monitor their children's snacks during school.

Second, the paper demonstrates long-lasting effects of the intervention, as the improvements in diet are on average *fully present* four months after the end of the intervention and removal of the incentives. This is rather 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, in the domain of healthy choices, 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.

Finally, this paper also contributes from a methodological perspective. Previous studies providing nutritional information have done so through labels that either simply manipulate attractiveness (Morizet *et al.*, 2012; Pelchat and Pliner, 1995) or convey 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 color 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 color-code interventions have had longer-term effects. Our intervention departs from those in the literature through the use of numerical grades identical to those assigned for school courses. Unlike caloric information, children had copious experience with these numerical grades. Unlike color labels, these numerical grades gave children a yardstick by which they could impress their parents as well as compete with one another. Just like children (and adults) compete in exercise, they turn out to compete in heathy eating when provided a means of keeping score.

Our intervention, particularly the treatment involving parents, points to a low-cost and highly-effective means of 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.

2. Previous literature

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 children's food choices. Second, this paper also connects to research on the effect of providing food information on healthy decisions. The literature has investigated two strategies that parents use to influence their children's behavior: restriction and pressure. Restriction has been found to have negative consequences on 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. We know of no studies identifying restriction as 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 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 intake was higher in the gentle condition.

Researchers also have explored the effectiveness of providing 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. Although they had short-term effects on consumption when displayed, calorie labels had no lasting effect on consumption once removed.

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. Choices in the intervention site were healthier than those in the comparison sites for some items, while for others the 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 was 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.[2] These incentives have been shown to have an effect in the short-term, with little exploration of long-term effects. For example, Just and Price (2013) conducted a field experiment at fifteen elementary schools in Utah. They rewarded students for eating a serving of fruit or vegetables per day in various different ways and found a 27% increase in the fraction of children eating at least one serving of fruit or vegetables when any incentive was offered. But they 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. Both schemes provided participants with a sticker for choosing a fruit or vegetable. In the piece-rate treatment, subjects who collected four stickers over the week received a prize. In the tournament treatment, only the child with the largest number of stickers

[2] Material incentives have been used to promote healthy behaviors other than the one addressed in this paper. For example, Charness and Gneezy (2009) were the first to 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 during the intervention (when the incentives were in play), but not afterwards. Of course, monetary incentives are difficult to roll out on a large scale and daunting to maintain in the long run.

won the prize. Results show that there was little effect of the piece-rate scheme on choices yet 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). They ran 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 proved effective: the proportion of children choosing the healthy snack increased from 17% in the baseline to nearly 80% with monetary incentives. The authors find some evidence of short-run post-intervention effects in the form of behavior one week after removal of the incentives but do not provide any evidence on longer-term efficacy. Compared to List and Samek (2015), our paper also studies the role of parents as potential change agents.

Loewenstein *et al*. (2016) conducted a field experiment in elementary schools in Utah. During a three- or five-week 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 authors find a strong positive effect of the reward on diet: the percentage of children who ate at least one serving of fruits or vegetables increased from 38-40% to 76-80%. However, two months after removal of the incentives, this percentage dropped dramatically (to 48-54%), although to a level above that in the baseline condition. Roughly 70% of the increase over the baseline dissipated over the two months following the intervention. In our paper, we are able to study the potential long-term effects of non-monetary interventions and how their potential persistence depends on the different non-monetary incentives.

Finally, Belot *et al.* (2019) conducted a field experiment to examine the malleability of dietary habits. 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. Children in both treatments reduced their body mass index and sugar intake compared to children in the control group. 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. Experimental design and procedures

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.[3] 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.[4]
- *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 (Spanish school 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.[5] Children were just shown the grades associated with each tray and were told that these grades represented the nutritional value of the items in 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.

[3] See Appendix A.1 for the nutritionist's justification of the allocation of food items to trays, and Appendix A.2 for a summary of the food nutritional information.
[4] A summary of the nutritionist's talk is reported in Appendix B.
[5] See Appendix C for the detailed composition of the trays and the corresponding grade associated with each one.

Children received exactly the same information as in *GT*. In addition (and before selecting the items), children were also aware of the fact that their parents would receive a weekly 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.

3.2 Procedures

The experiment was conducted in 12 Spanish elementary 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 amongst the children.[6] 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.[7] Each participating 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.[8] One class was randomly selected from each school to participate in the experiment. The total sample of 282 students between the ages of 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

[6] In 2018, 79,000 schools across Europe involving over 30 million children participated in this program.
[7] 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.
[8] 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.

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. In order to minimize interaction with the other participants, children made their decisions alone. To mitigate social-desirability effects, the bags used to keep the food were transparent so the researcher could see which snacks the children picked, without being directly involved during the decision. The intervention was always conducted right before the main school break at 11 a.m., when children go outside to get exercise; they also eat a snack during the break to tide them over until lunch, which—in Spain—starts at 2:00 - 3:00 p.m.[9] To make the food decision salient, in every treatment the parents were instructed not to prepare any snacks for their children for the days of the intervention.[10] 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.[11]

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.[12] 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

[9] Note that only a fraction of Spanish students (22%) has lunch at the school canteen. For those who eat at the school, the schools provide a menu, and children do not have the chance of choosing their food during lunch.
[10] 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.
[11] We did not find any differences across treatments in the amount of food saved for later consumption.
[12] The procedures to collect the data on subsequent days were the same as on the first day, i.e., participants would be located in different rooms and decisions would be made privately.

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 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 (see Appendix D.1 to D.3 for the complete questionnaires). Regarding the socioeconomic background of the students, we utilized a questionnaire administered by the Spanish Government and filled in 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. For the self-control indicators, following Tsukayama *et al.* (2013), children answered questions related to their level of interpersonal and schoolwork impulsivity. 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). 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 initial dietary habits or socio-economic characteristics across the subjects assigned to the four treatment arms, indicating that randomization was

successful.[13] Nonetheless, some of the regression models reported in Section 5.4 incorporate covariates in order to account for any potentially remaining imbalance (Raab and Butcher, 2005).

Finally, four months after the end of the intervention and during the next academic year, subjects participated in a *surprise* session, which was conducted as a one-shot *Baseline* for everyone and which was unannounced.[14] The aim of this was to study whether the effect of the different incentives remained after some time and once incentives had been removed.

4. Theoretical predictions and hypotheses

Let $X = \{x \in N^L : \sum_{l=1}^{L} x_l = k\}$ for some integer $k > 1$ be the set of feasible combinations of the L snacks (with non-negative integer values of each snack) available to a student. Let $\ominus \subset R^L$ be a set of parameter values that describe the healthiness of the L available snacks. The student's utility-from-consumption function $v: X \times \ominus \times R \rightarrow R$ describes the student's direct preferences over snack bundles. We assume that a student i who chooses bundle $x^i \in X$ at any given date derives *direct utility*:

$$v(x^i, \theta, \alpha) = t(x^i) + \alpha \sum_{l=1}^{L} \theta_l x_l^i$$

where $t(x^i)$ denotes the student's enjoyment of the *taste* of the bundle $x^i = (x_1^i, \ldots, x_L^i)$, $h(x^i, \theta) = \sum_{l=1}^{L} \theta_l x_l^i$ represents the *healthiness* of the bundle, and $\alpha > 0$ measures the student's concern for health.[15] For example, θ_l might be equal to minus the calories of snack l, or it might be equal to an indicator function that takes on the value of one if and only if l belongs to the healthiest quintile of snacks. Whereas the health value $h(x^i, \theta)$ is assumed to be additively

[13] 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.

[14] 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.

[15] It would be straightforward to generalize the model to have utility from consumption be an increasing but not necessarily additive, function of $t(x)$ and $h(x,\theta)$.

separable in the quantities of the different snacks, x_l^i, the taste utility $t(x^i)$ is not; this allows the student to care about the composition of his bundle (e.g., he might want to diversify away from the four items of the single snack he finds tastiest). The student knows his preference parameter α but may be uncertain about health consequences $\theta = (\theta_1, \ldots . \theta_L)$.

In the *Baseline*, children make choices without knowing θ. They solve:

$$\max_{x \in X} E[v(x, \theta, \alpha)] \Leftrightarrow \max_{x \in X} \left\{ t(x) + \alpha \sum_{l=1}^{L} E[\theta_l] x_l \right\} \Leftrightarrow \max_{x \in X} v(x, \alpha\, E[\theta], 1),$$

which means that a student who maximizes the expectation over θ of utility simply maximizes his direct utility given the expected value of θ. Let $x^*(\theta, \alpha)$ denote the student's optimal choice as a function of the parameters θ and α; for simplicity, we take this optimal choice to be single-valued.[16]

Fact 1: *For* each snack l, the optimal consumption level $x_l^*(\theta_l, \theta_{-l}, \alpha)$ is a non-decreasing function of θ_l.[17]

Ceteris paribus, a student will consume more of snack *l* as its healthiness increases.

Proof. Let $x_{-l}^*(\theta_{-l}, \alpha; x_l)$ denote the optimal bundle when consumption of snack l is fixed to be x_l. Note that neither $x_{-l}^*(\theta_{-l}, \alpha; x_l)$ nor $t\big(x_l, x_{-l}^*(\theta_{-l}, \alpha; x_l)\big)$, the taste utility from the bundle $(x_l, x_{-l}^*(\theta_{-l}, \alpha; x_l))$, depends upon θ_l. The student chooses $x_l \in \{0,1, \ldots, k\}$ to maximize

$$t\big(x_l, x_{-l}^*(\theta_{-l}, \alpha; x_l)\big) + \alpha\theta_{-l} \cdot x_{-l}^*(\theta_{-l}, \alpha; x_l) + \alpha\theta_l x_{l,}$$

which is supermodular in (x_l, θ_l). Since the feasible set is a lattice, the result follows from Topkis's (1978) Monotonicity Theorem.

Fact 2: The healthiness of the optimal choice $h(x^*(\theta, \alpha), \theta) = \sum_{l=1}^{L} \theta_l\, x_l^*(\theta;\ \alpha)$ is a non-decreasing function of α.

Proof. Define $\hat{t}(h) = \max_{x \in X}\{t(x) : \theta \cdot x = h\}$, the highest taste utility from a bundle of

[16] Although the optimal choice will be multi-valued for a non-generic set of parameter values, we abstract from this detail to simplify exposition.
[17] Throughout, the subscript $-l$ refers to snacks $j \neq l$.

healthiness level h. The maximization problem can be expressed as maximizing $\hat{t}(h) + \alpha h$ over the set $\{h \in R: h = \theta \cdot x, x \in X\}$. Because the domain is a lattice, and the objective function is supermodular in (h, α), the result again follows from Topkis's Theorem.

The *Nutritionist Treatment* makes nutrition salient to the children, which we conceptualize as increasing the weight α that the children assign to health when choosing their snacks. This gives us:

Prediction 1: The *Nutritionist Treatment* gives healthier choices than the *Baseline*.

The *Grades Treatment* makes two conceptually distinct changes to the *Baseline*. First, it reveals information about θ. Second, it gives students a yardstick by which to compete. We examine each of these effects in turn.

Suppose that students know the nutrition of all foods other than Snack 1. Fact 1 implies that a student who learns good news about Snack 1 consumes no less of it after learning θ, whereas a student who learns bad news about Snack 1 consumes no more of it. Revealing nutritional information about Snack 1 has an ambiguous effect on the overall healthiness of the student's chosen bundle. To see how better information about health may worsen health outcomes, consider a case in which a student avoids tasty Snack 1 in *Baseline*, believing it to be very unhealthy (i.e., his $E[\theta_1]$ is low). If Snack 1 turns out to be healthier than predicted, $\theta_1 > E[\theta_1]$, he may substitute away from healthier foods into Snack 1, decreasing the overall healthiness of his chosen bundle.

In certain cases, however, the effect is unambiguous.

Fact 3: If for each good l, (i) $E[\theta_l] = \bar{\theta}$, or if (ii) α is sufficiently large, then $h(x^*(\theta, \alpha), \theta) \geq h(x^*(E[\theta], \alpha), \theta)$.

Proof. Under assumption (i), the student's problem becomes:

$$\max_{x \in X} \left\{ t(x) + \alpha\bar{\theta} \sum_{l=1}^{L} x_l \right\} \Leftrightarrow \max_{x \in X} \{ t(x) + \alpha\bar{\theta}k \} \Leftrightarrow \max_{x \in X} t(x) \Leftrightarrow \max_{x \in X} v(x, \theta; 0).$$

Fact 2 implies that healthiness is lower than $h(x^*(\theta;\ \alpha), \theta)$. Under (ii), the student maximizes $E[\theta \cdot x]$. Blackwell's Theorem (1953) implies that having better information

(learning θ) does not decrease the value of the objective function and therefore healthiness.

In the *Grades Treatment*, grades not only provide students with information about the healthiness of the different foods, but they also give a yardstick by which students can compete with one another. To capture the competitive angle of the *Grades treatment*, we enrich students' preferences to incorporate some concern about their grades relative to those of their peers. Let $r\big(h(x^i,\theta),(h(x^j,\theta))_{j\neq i}\big)$ measure that part of *i*'s utility that derives from a comparison of the healthiness of *i*'s snacks relative to the healthiness of his peers' snacks. We assume that *r* increases in its first argument $h\big(x^i,\theta\big)$ and is, for each $j\neq i$, supermodular in $h\big(x^i,\theta\big)$ and $h\big(x^j,\theta\big)$: healthy eating by student *i* is a *strategic complement* (Bulow *et al.*, 1985) to healthy eating by any other student. Taking other students' snack choices $(x^j)_{\,j\neq i}$ as given, student *i* chooses

$$\max_{x^i\in X}\{v\big(x^i,\theta,\alpha\big)+\beta r(\theta\cdot x^i,(\theta\cdot x^j)_{\,j\neq i})\},$$

for $\beta>0$. Let $u(x,\theta,\alpha,\beta)$ denote the objective function in this maximization problem, and $x^*(\theta,\alpha,\beta)$ its maximizer.

> **Fact 4:** The healthiness of the optimal choice $h(x^*(\theta,\alpha,\beta),\theta)$ is non-decreasing in β.
> Because the proof of Fact 4 follows the same lines as that of Fact 2, we omit it.

When students have very accurate initial beliefs about nutrition, then the *Grades Treatment* does not change their beliefs about θ; we show in the next section that students in our experiment do indeed have rather accurate beliefs about the nutrition of the various snacks without seeing any grades. Consequently, moving from the *Baseline* to the *Grades Treatment* corresponds to an increase in β, which Fact 4 implies increases healthiness.

> **Prediction 2:** The *Grades Treatment* gives healthier choices than the *Baseline*.

Finally, we conceptualize the *Parents Treatment* as introducing the same β parameter as the *Grades Treatment* while also increasing α relative to the *Grades Treatment*. The idea behind this conceptualization is that parents value their children's healthy eating more than the children

do themselves; the *Parents Treatment* makes children weight their parents' welfare more heavily in their decision-making, which leads them to up-weight healthiness.[18] This gives:

Prediction 3: The *Parents Treatment* gives healthier choices than the *Grades Treatment*.

5. Results

This section is structured as follows. We begin by comparing the average behavior of children in each treatment using non-parametric tests and regression analysis. We then examine the dynamics and evolution of decisions over time during the intervention period and the questions of persistence in the longer run.

5. 1. Children's aggregate 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, which we define as items assigned grades of 7.5 or above.[19]

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

Treatment	Observations	Subjects	Average grade	Proportion of healthy choices
Baseline	398	73	4.77 (2.09)	0.36 (0.26)
NT	411	71	5.46 (1.99)	0.45 (0.27)
GT	373	65	5.63 (1.85)	0.47 (0.25)
PT	429	73	7.88 (1.61)	0.74 (0.24)

Notes: Standard deviations are reported in parentheses.

As seen in Table 1, the average grades in *NT* (5.46) and in *GT* (5.63) are larger than in *Baseline* (4.77). Furthermore, as shown in Table E.5 of the Appendix, the grades in every school

[18] Formally, suppose that a student in *GT* maximizes $u(x,\theta,\alpha,\beta)$, whereas his parents have preferences $u(x,\theta,\alpha',\beta)$, for $\alpha'>\alpha$. When the student in *PT* gives his parents' utility the weight $\delta>0$ by maximizing $u(x,\theta,\alpha,\beta)+\delta u(x,\theta,\alpha',\beta)$, this is equivalent to maximizing $u(x,\theta,\alpha'',\beta)$, for $\alpha''=\frac{\alpha+\delta\alpha'}{1+\delta}>\alpha$.

[19] 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 the Appendix).

of *NT* and *GT* are higher than in any school of the *Baseline*, so we have statistical significance when comparing *Baseline* to either *NT* or *GT* non-parametrically ($p = 0.050$ for the one-tailed Mann-Whitney tests of our directional hypotheses, using school-level data as independent observations).[20,21] There is no significant difference between the averages in *NT* and *GT*.[22]

So simply providing information to children, either by having grades assigned to the food or by having a nutritionist explain to children the benefits of eating in a salubrious manner does seem to increase the average rate of healthy food choices over the course of the intervention. Moreover, the two ways of providing information seem to have very similar effects on the average health quality chosen. These results support both *Prediction 1* and *Prediction 2*.

In order to study which of the two factors proposed in the theoretical model drives the results in *GT*, we conducted a questionnaire to capture the accuracy of children's knowledge regarding foods' healthiness. Fifty-eight boys and girls aged 9-10 (different from those who participated in the experiment) completed this questionnaire. They were asked to rank the food trays used in the experiment and grade them according to their nutritional value (see Appendix D.4 for the questionnaire). Results show that 80% of the children had very good awareness of the nutritional value of the baskets (making at most one mistake in their ranking), with 48% of participants making no mistakes. This is consistent with previous literature (Nguyen, 2008; Varela and Salvador, 2014) showing that children have excellent awareness about the healthiness of the food. These results support the idea that the change in children's behavior in *GT* relative to *Baseline* mainly derives from competition.

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

[20] 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 generally 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.

[21] 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 rank-sum 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*).

[22] 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 the same using *t*-tests ($t = 0.241$, *p*-value = 0.411, one-sided) or permutation tests ($p = 0.393$, one-sided).

possible significance level for each pairwise comparison to the *PT* treatment (p = 0.05 in all cases).[23] These results support *Prediction* 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.[24]

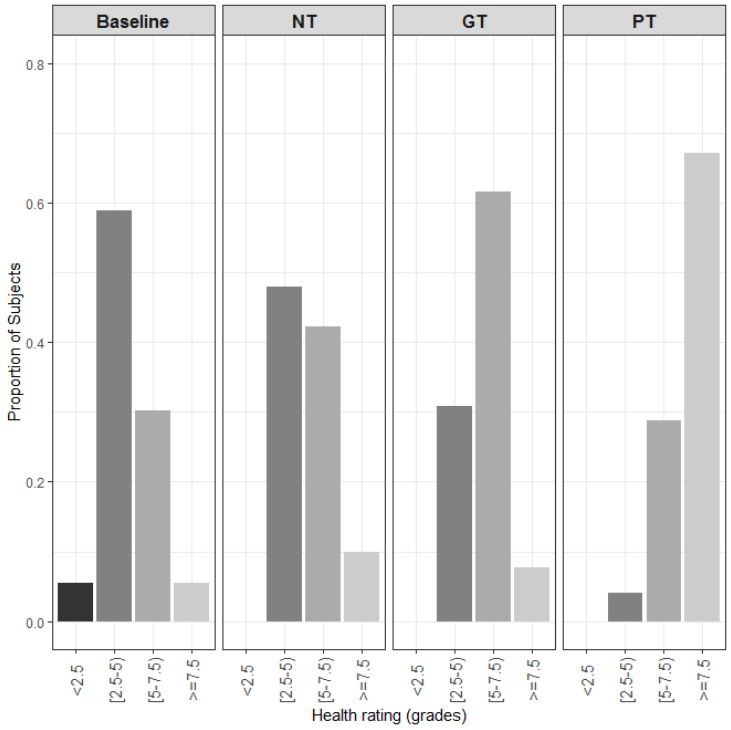
Turning to the proportion of healthy choices, results are very similar to those using the average grades. As shown in Table 1, the percentage of healthy items chosen in the *Baseline* is 36%. This percentage increases in *NT* (45%) and *GT* (46%) and more than doubles in *PT* (74%).

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 parents (in PT) has a very strong effect on subjects' decisions, largely increasing the consumption of healthy food items*.

[23] As in footnote 22, we provide the results of t-tests and permutation tests for robustness: 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.*

[24] As seen in Table E.5 in the Appendix, average grades are also higher for every school in *PT* vis-à-vis any school in the other three treatments.

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



We now explore the distribution of children's individual behavior to gauge the heterogeneity in treatment effects from providing grades to parents. To do so, we compute the average grade of each individual and then allocate it to one of the following average-grade categories: *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 these four categories.

Figure 1 shows that the proportion of subjects in each of the four categories differs considerably across treatments. Only *Baseline* has (a few) subjects in the worst category (with average grades at or below 2.5), and it has also the largest share of subjects in the second-worst category. The latter fraction is clearly decreasing across treatments from left to right. Overall, the distribution is significantly different between *Baseline* and *NT* (one-tailed Kolmogorov-Smirnov test: $\chi^2_2 = 3.92$, $p = 0.070$), and also between *Baseline* and *GT* ($\chi^2_2 = 37.89$, $p < 0.001$). Most importantly, there is a vast difference in the distributions across categories between *PT* and any other treatment, with test statistics of $\chi^2_2 = 55.48$, $p = 0.000$ for the comparison *PT* versus *Baseline*, $\chi^2_2 = 51.57$, $p = 0.000$ for the comparison *PT* vs. *GT*, and $\chi^2_2 = 47.88$, $p = 0.000$ for the comparison *PT* vs. *NT*. In particular, 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; which significantly exceeds the share in *Baseline* (5.5%), *GT* (12.20%) and *NT* (9.85%).[25] 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.[26]

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

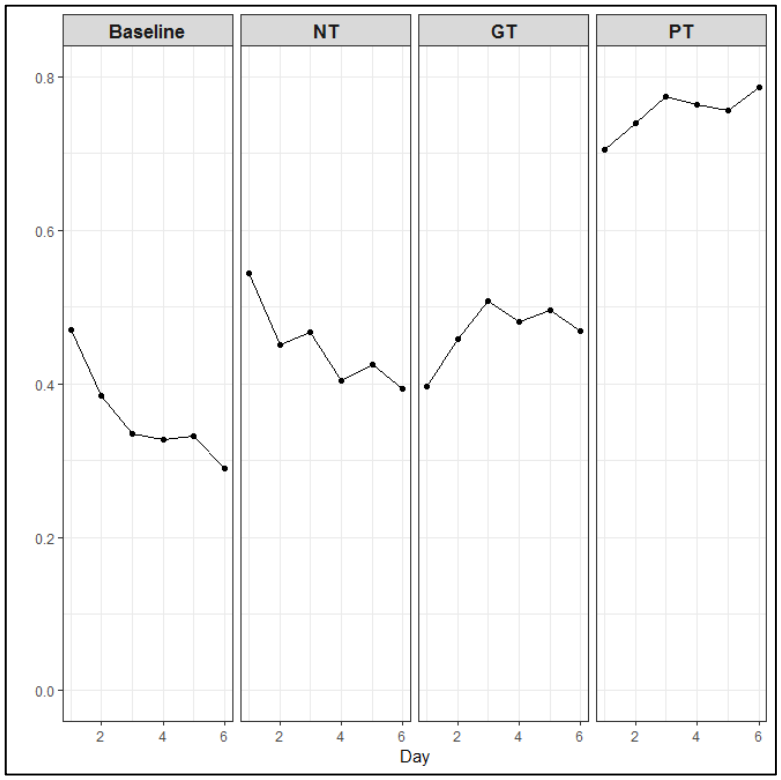
[25] All differences are statistically significant at the 5% level with (one-tailed test) Mann-Whitney tests on school-level data using *PT* as the reference group. These differences with respect to *PT* are significant at the 1% level using a (one-tailed) *t*-test. Cluster (school)-adjusted chi-square tests for individual-level data yield similar conclusions: χ^2 = 4.864, p = 0.027 versus *Baseline;* χ^2 = 4.909, p = 0.027 versus *GT*; and χ^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 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*.

[26] This conclusion is also confirmed when we look only at school level data, as we show them in Table E.5 in the Appendix.

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 (as in Table 1) Figure 2 considers as healthy choices items that were assigned a grade of 7.5 or 10.[27]

Figure 2. Percentage of healthy choices over time



Because we did not formulate any *ex-ante* predictions about dynamics, none of our predictions in Section 4 addresses dynamics; indeed, the formal model in Section 4 is static. However, this static model leads naturally to certain hypotheses about the dynamics of behavior in *GT* and *PT*. If we look at how decisions evolve over time, some patterns might emerge that would help us to better understand participants' behavior. Because these hypotheses were first articulated after the experiment, we state them only informally here.

[27] Figures E.1 and E.2 in the Appendix replicate that analysis using children's average grades and a stricter definition of healthy choices (i.e., defining as healthy choices those items that were assigned a grade of 10), respectively. The main patterns emerging from Figure 2 remain unchanged.

In both *GT* and *PT*, students receive information that allows them to compete with their peers or to please their parents; for students to play mutual best responses from Day 1, they must anticipate their own concern for relative grades as well as correctly predict their peers' choices prior to choosing their first bundle of snacks. Schoolchildren likely err in both respects. Suppose that students in Day 1 act as if $\beta = 0$, namely they neglect grade competition at the time of choosing their first bundle of snacks, when they have yet to learn anything about any peer's grades. Suppose that students on day $t > 1$ use their true β but naively forecast that their peers will earn the same grades that these peers earned in period t-1. Then we would expect students in the *Grades Treatment* to eat more healthily over time.[28] Indeed, teachers informed us that children in class were comparing grades and talking about how to improve the next day, so it appears that competition played a role here, as it did in Belot *et al.* (2016).[29]

The *Parents Treatment* presents a second channel through which one might expect healthiness to increase over time. Suppose children react to admonishment by their parents. Since parents view their children's grades for the first time after Day 2, the first opportunity for children to discuss grades with their parents arises between Day 2 and Day 3. Then some children, chastened by their initial parent-child nutrition chats, would attend more to health starting on Day 3. However, as observed earlier, it could also be that many students actually anticipate the reaction of their parents and preempt any unpleasant conversation by eating healthy snacks from the beginning. Figure 2 shows a positive general trend in *GT* and *PT*, with the improvement mainly from Day 1 to Day 3. It also shows that the percentage of healthy choices in the first period is 70.52% in *PT*, which is much larger than in *Baseline* (46.68%), *NT* (54.37%), and *GT* (39.68%).[30] These results indicate that even just knowing that their parents

[28] Fact 4 implies that they will eat healthier foods on Day 2 than Day 1. Combining this result with the fact that grades are strategic complements implies that children will eat healthier on Day 3 than Day 2. Iterating this argument leads to the conclusion that healthy eating increases over time in *GT*.

[29] Other examples of the effect of competition include Bornstein *et al*. (2002), who find that competition between groups increased contributions in the minimum-effort game, Fershtman *et al*. (2012), who find that just having the notion of competition in the air led to social preferences vanishing (albeit Houser and Schunk, 2009, find such an effect of competition on social preferences only for boys), and Charness and Holder (2019), who show that competition between groups to have their charitable 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.

[30] Note that on the first day of the experiment children do not know anything regarding what will happen in the experiment after they make their decision. Hence, while eating patterns may be correlated within schools, we can ignore any dynamic effects. We thus conduct cluster (school)-adjusted chi-square tests for individual-level data from this first day. Differences are statistically significant for the comparison *PT* versus *Baseline* ($\chi^2 = 0.445$, $p = 0.063$),

will view this information is sufficient to have a large effect on the food choices made by the students. Finally, the lack of explicit incentives seems to cause a reduction in the proportion of healthy items, which leads to the uniform negative trend observed in *Baseline* and *NT*.[31]

5.3. Regression analysis

Table 2 reports parameter estimates from probit models in which the dependent variable is a dummy taking the value 1 if subject *i* made a "healthy choice" (i.e., subject i's choice had an average grade of 7.5 or higher) in period *t*, and 0 otherwise. All specifications use cluster-robust standard errors at the school level to account for heteroskedasticity and intra-school correlation.[32]

Since the number of schools in our sample is quite small, conventional inference methods may be unreliable: 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 Wald hypothesis tests computed according to Kline and Santos' (2012) score bootstrap procedure for clustered data, which adapts the wild bootstrap-*t* procedure developed by Cameron *et al*. (2008) to maximum likelihood estimators. This approach is useful for small sample sizes like ours because it does not rely on asymptotic approximations, and it has been shown to perform well with very small numbers of clusters (Kline and Santos, 2012). [33]

The explanatory variables in the first column of Table 2 are the indicators for each treatment (*NT*, *GT*, and *PT*, taking *Baseline* as the benchmark). Subjects' probability of making a healthy choice increases in all three treatments compared to *Baseline*, even after accounting for the small

---

for *PT* versus *NT* ($\chi^2 = 2.746$, $p = 0.097$), and for *PT* versus *GT* ($\chi^2 = 4.140$, $p = 0.042$). Cluster-adjusted, one-tailed *t*-tests give $t = 3.027$, $p = 0.019$ versus *Baseline*; $t = 2.162$, $p = 0.048$ versus *NT*, and $t = 4.089$, $p = 0.007$ versus *GT*. Tests with school-level data also indicate significance at the 5% level.

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

[32] Since all the explanatory variables included in the models presented in Table 2 are time-invariant, these specifications do not include individual, school or period (day) fixed effects. For robustness, Table E.8 in the Appendix reports estimates from multi-level probit 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 2.

[33] Tables E.9 - E.10 in the Appendix report estimates from linear regression models in which the dependent variable is subject *i*'s grade in period *t.* Table E.9 resorts to Cameron *et al.*'s (2008) wild bootstrap-*t* procedure for clustered data in order to account for the small number of schools in our sample, while Table E.10 presents estimates from multi-level regression models. The key substantive results from these additional specifications are similar to those in Table 2.

Table 2. Probit regression on the probability of choosing healthy food items

	(1)	(2)	(3)
Constant	-1.34***	-1.69***	-1.66***
	(0.08)	(0.40)	(0.48)
Wild bootstrap-t *p*-value	(0.00)	(0.00)	(0.00)
NT	0.27**	0.33**	0.33**
	(0.12)	(0.13)	(0.16)
Wild bootstrap-t *p*-value	(0.04)	(0.05)	(0.05)
GT	0.21*	0.33**	0.44**
	(0.12)	(0.15)	(0.17)
Wild bootstrap-t *p*-value	(0.06)	(0.05)	(0.03)
PT	1.39***	1.37***	1.40***
	(0.11)	(0.12)	(0.15)
Wild bootstrap-t *p*-value	(0.02)	(0.01)	(0.00)
Male		0.10	0.16*
		(0.08)	(0.09)
Wild bootstrap-t *p*-value		(0.55)	(0.27)
Average school grade		0.11***	0.09**
		(0.04)	(0.04)
Wild bootstrap-t *p*-value		(0.03)	(0.08)
Personal Impulsivity		-0.13*	-0.15**
		(0.07)	(0.07)
Wild bootstrap-t *p*-value		(0.09)	(0.09)
School Impulsivity		-0.11	-0.07
		(0.08)	(0.09)
Wild bootstrap-t *p*-value		(0.19)	(0.24)
Parents hold University degree			0.03
			(0.11)
Wild bootstrap-t *p*-value			(0.32)
Household Income			0.05
			(0.11)
Wild bootstrap-t *p*-value			(0.47)
# Observations	1,611	1,347	1,129
Pseudo-R^2	0.16	0.17	0.17

Notes: Maximum likelihood estimation. 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 method that accounts 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.

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' probability of choosing healthy food is significantly larger in *PT* than in *GT* (χ^2 = 17.8, p = 0.000; test for equality of coefficients) and *NT* (χ^2 = 22.77, p = 0.000). There is no statistically-significant difference between the effect of the grades-only treatment and the effect of the nutritionist, though (χ^2 = 0.50, p = 0.478; test for equality of coefficients between *GT* and *NT*).

The treatment effects remain robust if we incorporate additional controls in column (2): *Male*, a dummy for male subjects; *Average School Grade*, a measure of subjects' academic performance, and two measures for impulsivity levels. For this, we built two indexes that measure two domain-specific impulsivity levels (Tsukayama *et al.*, 2013): *Personal Impulsivity* and *School Impulsivity*, which we include as covariates. *Average School Grade* is significantly and positively associated with the probability that subjects make healthy food choices. Among the impulsivity measures, only *Personal Impulsivity* is (marginally) significantly correlated with the probability that subjects choose healthy food items.

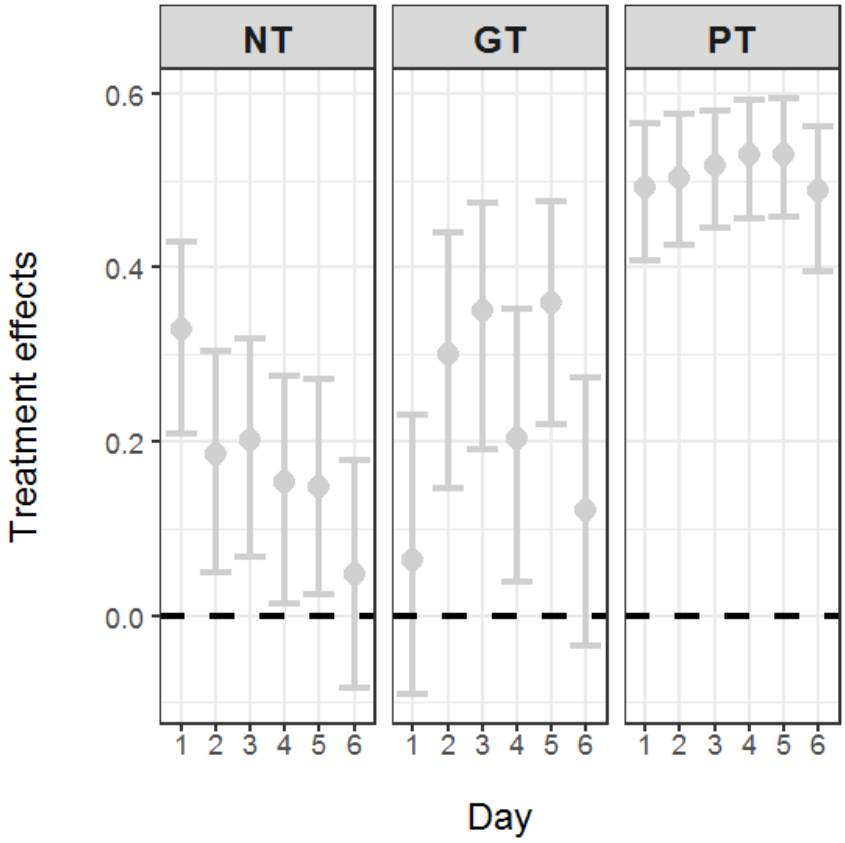
Column (3) replicates the analysis in column (2), but also takes into account the socio-economic level of a child's family. Specifically, *Parents hold University degree* is a binary variable that takes value 1 if at least one of the parents achieved graduate or postgraduate education and 0 otherwise; 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.[34,35] We observe that a higher educational level of parents does not significantly affect their child's grade, and neither does the household's economic level. The estimates for *PT*, *GT* and *NT* remain statistically significant after controlling for these additional covariates.

[34] 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).

[35] 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 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 remain significant across columns (1) to (3) – see also Tables E.8 - E.10 in the Appendix – mitigates this concern.

Next, we consider the evolution of the treatment effects over the course of the intervention, Figure 3 plots the differences in the probability that the average student chooses a healthy food item between the *Baseline* treatment and *GT*, *NT* and *PT* for each period (i.e., the daily treatment effects).[36]

Figure 3. Differences in healthy food choice rates in comparison to *Baseline*



Notes: For each day, the figure plots the difference in the probability of making healthy food choices between *NT* (left panel), *GT* (middle panel) and *PT* (right panel) and the *Baseline* treatment. Solid circles represent point estimates; vertical lines give the 95% confidence intervals, obtained by inverting the score bootstrap tests.

Results show that differences between *PT* and *Baseline* are positive and statistically significant throughout the experiment. The differences between *GT* and *Baseline* become

[36] The per-period effects in Figure 3 are based on the parameter estimates reported in Table E.11 in the Appendix, which include only the treatment variables interacted with period indicators as predictors. The confidence intervals in Figure 3 are obtained by inverting the score bootstrap tests, and are not necessarily symmetric (Roodman *et al.,* 2019). They are typically more conservative than "standard" confidence sets relying on asymptotic approximations.

significant from the second day until day 5. Finally, there is a significant difference between *Baseline* and *NT* already on the first day (when the nutritionist gives the talk). This initial difference gradually fades out over time, becoming insignificant only on day 6. These findings support the idea that the positive effect of the treatment dummies on students' healthy food choices are fairly stable.[37]

5.4. Effects after the removal of incentives – The surprise session four months later

A key issue is whether the effects of the intervention persist over time. While some previous work has found 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.

Table 3 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.[38] Even after removing the incentives and four months after the intervention period, the proportion of healthy choices in each of *NT* (47%), *GT* (53%) 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 vs. GT* and *PT vs. NT*).[39] These results indicate two main conclusions. First, there is a positive effect of *GT* on

[37] Similar findings are obtained using subjects' average grades as the outcome variable (see Table E.12 and Figure E.6 in the Appendix).

[38] The lower number of observations in the surprise sessions is mainly due to the fact that one school in *Baseline* (Santo Domingo II) and one school in *PT* (Gloria Fuertes II) decided not to participate in the surprise sessions. As we show in Table E.6 of the Appendix, removing these two schools does not lead to systematic covariance imbalances across treatment arms.

[39] 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). In fact, Table E.7 in the Appendix, which reports the

healthy choices while the effect of the incentive weakens over time in *NT*. Second, providing information to the parents has a very strong effect, largely persisting even four months after the incentive was removed.

Table 3. Average grades and healthy food choices in the surprise session, by initial treatment

Initial Treatment	Observations	Subjects	Average grade	Proportion of healthy choices
Baseline	33	33	5.04 (1.70)	0.41 (0.25)
NT	68	68	5.60 (1.77)	0.47 (0.26)
GT	65	65	5.94 (1.91)	0.53 (0.27)
PT	45	45	7.43 (1.65)	0.69 (0.21)

Notes: In the surprise session, no incentives were used. Standard deviations are reported in parentheses.

We see that children did not behave that differently in the surprise session than the average behavior during the intervention. 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 at the individual level (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).[40] 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 (without incentives) four months after the experiment was quite similar to their average behavior in the three-week intervention period when the different incentives were in place, showing a remarkable degree of persistence even in the absence of incentives.*

5.5 Consistency of individual choices as a mechanism for the longer-term effects

One substantive issue potentially worth considering is the underlying mechanism for why our intervention produced effects. One possibility is that our intervention leads children to learn to like eating a particular combination of food items. This learning might create a habit that

average grades and the proportion of healthy choices by school, shows that the proportion of healthy choices is higher in each of the schools in *PT* than in any school in the other three treatments.

[40] Conclusions remain the same when we look at the average grades, also when using *t*-tests and permutation tests.

persists over time, and prevail even following removal of incentives. Hence, we study the extent to which the intervention leads to consistent, reformed eating habits, and whether any such consistency remains after the removal of incentives.

Table 4 presents a measure of individual consistency within the range of Days 2-6 (column [1]), across that interval and Day 1 (column [2]), and across that interval and Day 7 (column [3]). We define a subject's behavior as consistent if at least 75% of her choices belong to a pre-determined pool of food items, which is defined for each individual as her six most common food choices from days 2 to 6. Then, we compute how many subjects make at least 75% of their choices (i.e., they pick at least three out of four items per day) from the aforementioned pool in each one of days 2 to 6. For example, the bottom cell in column [1] in Table 4 shows that 67.34% of all participants in *PT* consistently picked at least three out of four items belonging to their pre-determined pool on each of days 2 to 6. Note that if a subject failed to do so on just one day, then this subject is not included in the reported percentage. In column [2], we first use the pre-determined pool of food items previously established for days 2 to 6. Then, we compute the percentage of subjects who picked at least 75% of their items from the pre-determined pool also on day 1. For example, the second cell in column [2] shows that 28.57% of all participants in *Baseline* picked at least three items belonging to their pre-determined pool already on day 1. Finally, in column [3] of Table 4, we do the same analysis as in column [2], but for day 7 (the surprise session) instead of day 1.

Table 4: Percentage Individual Consistency, Days 2-6 and Comparisons

Treatment	[1] Days 2-6	[2] Day 1 vs. Days 2-6	[3] Day 7 vs. Days 2-6
Baseline	42.86	28.57	61.90
NT	38.46	42.30	65.38
GT	48.52	19.11	38.23
PT	67.34	34.69	63.26

Notes: Column [1] refers to the percentage internally consistent for days 2-6. Column [2] refers to the percentage of consistent food choices between day 1 and days 2-6, and column [3] compares day 7 to days 2-6.

Column [1] shows that a much higher percentage of people were consistent in their choices through days 2-6 in *PT* than in the other treatments. Here it seems clear that the students found a relatively set menu that they thought would please their parents and stayed with it.

Columns [2] and [3] together suggest that the children learned over time to eat a particular set of items. In the surprise session (day 7), the respective percentages of consistent choices are 62%, 65%, 38%, and 63% for *Baseline, NT, GT,* and *PT*, which almost doubled the consistency on day 1 (29%, 42%, 19%, and 35%). This suggests that children got used to a particular set of foods (potentially different for each child) between days 2 and 6 and that they continued choosing largely from this set even four months after the intervention and once the incentives were removed. So, it looks like children return to eat foods that they had learned to like rather than those that they initially chose.[41]

6. Conclusion

Poor diet and obesity have been linked to a variety of contemporary 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, however, 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 to 2011-2012.

We conduct a field intervention designed to evaluate the effectiveness of different means of influencing children's diets. Previous work has shown positive effects of contemporaneous material benefits for healthy eating (Belot *et al*., 2016), but at best limited enduring effects thereafter. Our design shows that non-material incentives can be effective for leading children to make healthier food choices at school. We align children's appraisal of food choices with their

[41] To check the robustness of our results, we performed the same analysis as the one proposed in Table 4 but using a threshold of 100% (instead of 75%) and computing the pre-determined pool of food items using 8 items instead of 6. Results are qualitative the same as those reported in Table 4. This is shown in Tables E.12, E.13, and E.14 in the Appendix.

appraisal of schoolwork by introducing a system in which food items are graded based on their nutritional value. This provides students a yardstick by which to compete over healthy eating.

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*.

In terms of why the intervention produced such long-lasting effects, we see that it is not only the grades that are about the same in the surprise session (especially considering that no incentives were present there) but the foods are also the same. This fits rather neatly with the seminal exercise intervention study in Charness and Gneezy (2009), even though the mechanism lies outside the formal model. There students learned that they actually enjoyed the feelings derived from consistent exercise. Here there may be a lesson for future design: If one can get children accustomed to healthy snacks, then they will crave those snacks in the future.

Of course, if parental oversight is so effective, one may wonder why so many parents fail to ensure healthy eating at home or fail to provide their children with healthy lunch packages for school. One possible explanation is that parents in our experiment are involved in an indirect way (they just receive and monitor weekly grades reports) and need not engage in costly daily negotiations with their children over healthy eating. Children may prize the autonomy from maximizing preferences that attend to both taste and health. This shared effort could be key in reducing the burden on parents, hence making the intervention more effective

Our approach involves little or no financial cost, requires only monitoring from parents or peers, and has proved highly effective in both the short and medium run. It is obvious that more research would be helpful, since our study is limited mainly to middle-class families in Spain. If policy-makers 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.

References

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.

Blackwell, D. (1953) "Equivalent Comparisons of Experiments" *Annals Mathematical Statistics* 24: 265-272.

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.

Bulow, J, Geanakoplos, J, and Klemperer, P. (1985) "Multimarket Oligopoly: Strategic substitutes and complements," *Journal of Political Economy* 93: 488-511.

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., and Gneezy, U. (2009) "Incentives to exercise," *Econometrica, 77*(3): 909-931.

Charness, G., and Holder, P. (2019) "Charity in the laboratory: Matching, competition, and group identity," *Management Science*, 65(3): 1398-1407.

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.

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.

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.

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.

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.

Global Burden of Disease (2016). *The Lancet*. Available at http://www.thelancet.com/gbd, accessed 02 January 2018.

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.

Haire-Joshu, D., and Tabak. R. (2016) "Preventing obesity across generations: Evidence for early life intervention," *American Review of Public Health,* 37: 253-271.
Hansen, B.B., and Bowers, J. (2008) "Covariate balance in simple, stratified and cluster comparative studies," *Statistical Science,* 23(2): 919-936.
Houser, D., Schunk, D. (2009) "Social environments with competitive pressure: Gender effects in the decisions of German schoolchildren," *Journal of Economic Psychology*, 30: 34-641.
Ibragimov, R, and Muller, U.K. (2010) "t-Statistic based correlation and heterogeneity robust inference," *Journal of Business & Economic Statistics,* 28(4): 453-468.
Jansen, E., Mulkens, 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., Mulkens, 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.
John, L., Loewenstein, G., Troxel, A., Norton, L., Fassbender, J., and Volpp, K. (2011) "Financial Incentives for Extended Weight Loss: A Randomized, Controlled Trial," *Journal of General Internal Medicine*, 26(6): 621-626
Just, D.R., and Price, J. (2013) "Using incentives to encourage healthy eating in children," *Journal of Human Resources,* 48(4): 855-872.
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.
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.
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.
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.
Majolo, B., and Marechal, L. (2017) "Between-group competition elicits within-group cooperation in children," *Nature Scientific Reports*, 7: 43277.
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.
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.
Nguyen, S. (2008) "Children's evaluative categories and inductive inferences within the domain of food," *Infant and Child Development*, 17: 285-299

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.
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.
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.
Raab, G.M., and Butcher, I. (2005) "Randomization inference for balanced cluster-randomized trials," *Clinical Trials,* 2(2): 130-140.
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.
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.
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*.
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.
Topkis, D. (1978). "Minimizing a Submodular Function on a Lattice," *Operations Research* 26(2): 305-321.
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*.
Varela, P., and Salvador, A. (2014) "Structured shorting using pictures as a way to study nutritional and hedonic perception in children," *Food Quality and Preference*, 37:27-34.
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.

World Health Organization (2009), *Global Health Risks: Mortality and Burden of Disease Attributable to Selected Major Risks*, World Health Organization, Geneva.

Online Appendix

Appendix A.1: 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 A.2. Food Nutritional information

Basket 1 (0 points)

	Orange Juice (100 ml)	Pinapple juice (100 ml)	Bollycao (per unit)	Phoskitos (per unit)	Candy (15 g)
Calories	41	23	223	170	206
Fat	0.1 g	0 g	8.4 g	7.8 g	1.2 g
of which Saturates	0 g	0 g	1.7 g	6 gr	0.4 gr
Carbohydrate	10.4 g	5.4 g	31.5 g	24 gr	37 gr
of which Sugars	10.4 g	5.4 g	18.3 g	17 g	17.5 g
Protein	0.6 g	0.3 g	4.80 g	1.8 g	1.5 gr

Basket 2 (2.5 points)

	Nutella sticks (35 g)	Cereals (100 g)	Sandwich (100 g)	Milkshake (100 ml)
Calories	176	385	300	65
Fat	8.4 g	2.5 g	13 g	1 g
of which Saturates	2.9 g	0.9 g	9 g	0.6 gr
Carbohydrate	22 g	85 g	31.5 g	24 gr
of which Sugars	14 g	35 g	30 g	12 g
Protein	2.5 g	5.7 g	14 g	2.8 g

Basket 3 (5 points)

	Flavored Yogurt (125 g)	Milk (100 g)	Turkey on seeds bread sandwich (100 g)	Babybell cheese (per unit =20 g)
Calories	100	27	340	62
Fat	2.5 g	0.3 g	3 g	4.8 g
of which Saturates	1.6 g	0.2 g	1.5 g	4 gr
Carbohydrate	15.5 g	3 g	48 g	0 gr
of which Sugars	15.5 g	3 g	3 g	0 g
Protein	4.4 g	3 g	15 g	4.5 g

Basket 4 (7.5 points)

	Natural Yogurt (125 g)	Peanuts (100 g)	Grapes (100 g)	Pear (100 g)
Calories	68	570	65	50
Fat	3.8 g	50 g	0.02 g	0.1 g
of which Saturates	2.6 g	7 g	0 g	0 gr
Carbohydrate	4.2 g	16 g	20 g	12 gr
of which Sugars	4.2 g	5 g	17 g	7 g
Protein	4.3 g	25 g	0.02 g	0.5 g

Basket 5 (10 points)

	Water (330 ml)	Serrano Ham on rye bread sandwich (100 g)	Nuts (100 g)	Almonds (100 g)	Banana (100 g)	Apple (100 g)
Calories	0	550	650	600	125	63
Fat	0 g	15 g	65 g	51 g	0.4 g	0.5 g
of which Saturates	0 g	6 g	6 g	4 gr	0.4 g	0.5 g
Carbohydrate	0 g	50 g	13 g	5 gr	32 g	13.5 g
of which Sugars	0 g	2 g	2.5 g	5 g	16 g	12.5 g
Protein	0 g	20 g	15 g	25 g	1.4 g	0.5 g

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:

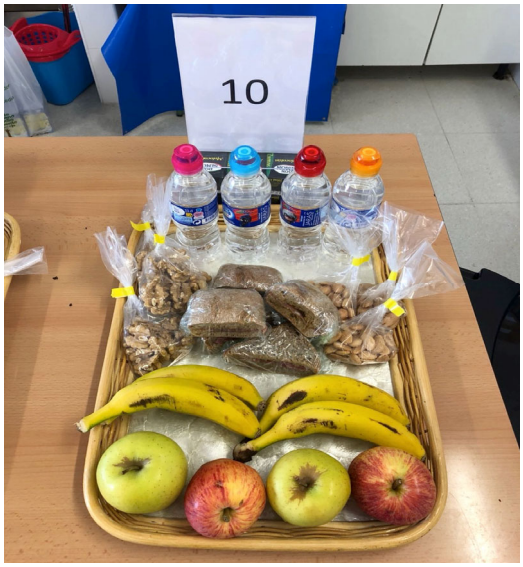
• In general, they produce or maintain an adequate state of health.

• 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


10

Appendix D: Post-experimental questionnaires

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				

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

D.3: Eating habits survey

WEEKLY CONSUMPTION					
	NEVER	**ONCE A WEEK**	**2/3 TIMES A WEEK**	**ALMOST EVERYDAY**	**ON A DAYLY 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					

D.4: Food knowledge questionnaire

Une con una flecha cada cesta de alimentos con la nota que tú crees que le corresponde.			
 	 	Cesta 1	10 (la bandeja más saludable)
 	 	Cesta 2	7,5
 	 	Cesta 3	5
 	 	Cesta 4	2,5
 	 	Cesta 5	0 (la bandeja menos saludable)

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	4.13	4.03	3.97	3.98
Eggs	2.52	2.71	2.80	2.74
Meat (beef, pork and chicken)	3.14	3.07	2.94	2.98
Seafood	2.14	2.22	1.98	2.08
Fruit and vegetables	2.91	2.98	2.79	2.91
Bread	2.91	2.95	2.91	2.99
Rice and Cereals	2.42	2.31	2.19	2.14
Cooking Oils	3.00	3.02	3.07	2.99
Confectionery	1.91	1.95	1.91	1.91
Other fats and sugars	2.14	2.01	2.10	1.94
Demographic and socio-economic characteristics				
Parents hold University degree[b]	0.03	0.05	0.06	0.08
Both parents are economically inactive[c]	0.04	0.03	0.03	0.02
Household Income[d]	26.22	35.04	28.69	30.61
Parents' involvement in children's school success[e]	3.53	3.56	3.47	3.46
Average school grade[f]	7.24	7.10	7.21	7.84
Frequency of child's school absenteeism[g]	0.05	0.04	0.09	0.02
Proportion of male subjects[c]	48.68	56.34	53.49	52.70

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

	NT – Baseline	GT - Baseline	PT – Baseline	GT-NT	GT-PT	NT-PT
Dietary habits (weekly consumption)						
Milk and dairy products	Z = -0.457 p = 0.455	Z = -0.748 p = 0.648	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 = 0.852 p = 0.394	Z = 1.434 p = 0.152	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 = -1.262 p = 0.207	Z = -1.045 p = 0.296	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 = 0.496 p = 0.620	Z = -1.031 p = 0.302	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 = 0.028 p = 0.978	Z = -0.838 p = 0.402	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 = -0.134 p = 0.894	Z = -0.01 p = 0.998	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 = 0.061 p = 0.951	Z = 0.011 p = 0.998	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 = -0.116 p = 0.907	Z = -0.405 p = 0.685	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 = -0.641 p = 0.522	Z = -0.985 p = 0.325	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 = -1.590 p = 0.112	Z = -0.189 p = 0.850	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 = 1.110, p = 0.266	Z = 0.401, p = 0.688	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 = -1.226, p = 0.220	Z= -1.017, p = 0.309	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 =1.803, p =0.070	Z = 0.450, p = 0.653	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 = 0.380, p = 0.704	Z = -0.778, p = 0.436	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 = -0.497, p = 0.619	Z = -0.099, p = 0.921	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 = -1.299, p = 0.194	Z = 0.841, p = 0.400	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 = 0.925 p = 0.355	Z = 0.501, p = 0.616	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	**$\chi2$= 10.700, p=0.151**	**$\chi2$= 5.230, p=0.514**	**$\chi2$= 6.670, p=0.464**	**$\chi2$= 10.101, p=0.185**	**$\chi2$= 4.581, p=0.710**	**$\chi2$= 5.941, p=0.546**

Note: Covariate balance 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
Nutritionist (NT)	*Constant*	2.31
		(4.37)
	Cluster-Adjusted p-values	(1.00)
	Milk and dairy products	-0.56
		(0.52)
	Cluster-Adjusted p-values	(0.34)
	Carbohydrates	-0.39
		(0.34)
	Cluster-Adjusted p-values	(0.45)
	Proteins	0.07
		(0.41)
	Cluster-Adjusted p-values	(0.41)
	Fruit and vegetables	0.20
		(0.70)
	Cluster-Adjusted p-values	(0.42)
	Fats and sugars	0.25
		(0.51)
	Cluster-Adjusted p-values	(0.88)
	Male	-0.56
		(0.72)
	Cluster-Adjusted p-values	(0.45)
	Parents hold University degree	-0.18
		(1.06)
	Cluster-Adjusted p-values	(0.32)
	Both parents are economically inactive	-1.26
		(0.87)
	Cluster-Adjusted p-values	(0.43)
	Household Income	0.79
		(0.45)
	Cluster-Adjusted p-values	(0.90)
	Involvement in children's school success	0.39
		(0.57)

	Cluster-Adjusted p-values	(0.58)
	Average school grade	-0.35 (0.27)
	Cluster-Adjusted p-values	(0.748)
Grades (GT)	*Constant*	-5.09 (4.21)
	Cluster-Adjusted p-values	(1.00)
	Milk and dairy products	-0.99 (0.50)
	Cluster-Adjusted p-values	(0.21)
	Carbohydrates	-0.15 (0.30)
	Cluster-Adjusted p-values	(0.86)
	Proteins	0.77 (0.38)
	Cluster-Adjusted p-values	(0.64)
	Fruit and vegetables	-0.01 (0.68)
	Cluster-Adjusted p-values	(0.43)
	Fats and sugars	-0.13 (0.50)
	Cluster-Adjusted p-values	(0.79)
	Male	-0.65 (0.66)
	Cluster-Adjusted p-values	(0.13)
	Parents hold University degree	0.07 (1.10)
	Cluster-Adjusted p-values	(0.31)
	Both parents are economically inactive	1.32 (0.82)
	Cluster-Adjusted p-values	(0.63)
	Household Income	0.68 (0.43)
	Cluster-Adjusted p-values	(0.34)
	Involvement in children's school	0.66

	success	(0.53)
	Cluster-Adjusted p-values	(0.97)
	Average school grade	0.26 (0.26)
	Cluster-Adjusted p-values	(0.23)
Parents (PT)	*Constant*	-1.33 (4.06)
	Cluster-Adjusted p-values	(0.62)
	Milk and dairy products	-0.81 (0.49)
	Cluster-Adjusted p-values	(0.38)
	Carbohydrates	-0.39 (0.30)
	Cluster-Adjusted p-values	(0.38)
	Proteins	0.40 (0.37)
	Cluster-Adjusted p-values	(0.51)
	Fruit and vegetables	0.34 (0.66)
	Cluster-Adjusted p-values	(0.57)
	Fats and sugars	0.03 (0.46)
	Cluster-Adjusted p-values	(0.59)
	Male	-0.46 (0.64)
	Cluster-Adjusted p-values	(0.39)
	Parents hold University degree	0.18 (0.98)
	Cluster-Adjusted p-values	(0.41)
	Both parents are economically inactive	0.24 (0.77)
	Cluster-Adjusted p-values	(0.38)
	Household Income	1.272 (0.41)
	Cluster-Adjusted p-values	(0.50)

	Involvement in children's school success	-0.05 (0.51)
	Cluster-Adjusted p-values	(0.12)
	Average school grade	-0.05 (0.25)
	Cluster-Adjusted p-values	(0.23)

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,* and *Both parents are economically inactive* 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.

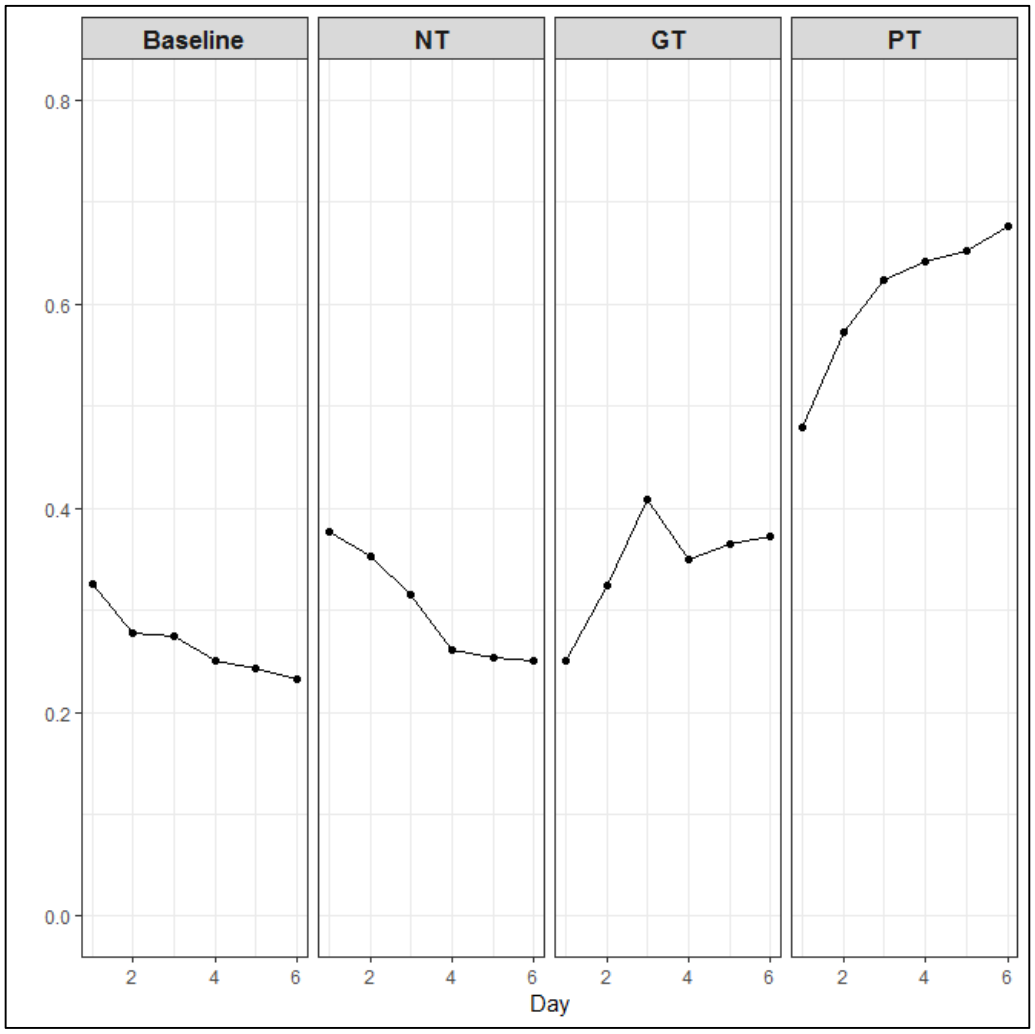
Table E.4. Proportion of healthy choices and differences across treatments using alternative definitions of "healthy choice"

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 *Baseline*	Proportion of healthy choices	Mann-Whitney tests vis-à-vis *Baseline*
Baseline	0.27	-	0.49	-
NT	0.30	$p = 0.200$	0.58	$p = 0.050$
GT	0.34	$p = 0.050$	0.59	$p = 0.050$
PT	0.60	$p = 0.050$	0.85	$p = 0.050$

Table E.5. Average grades and proportion of healthy food choices, by school

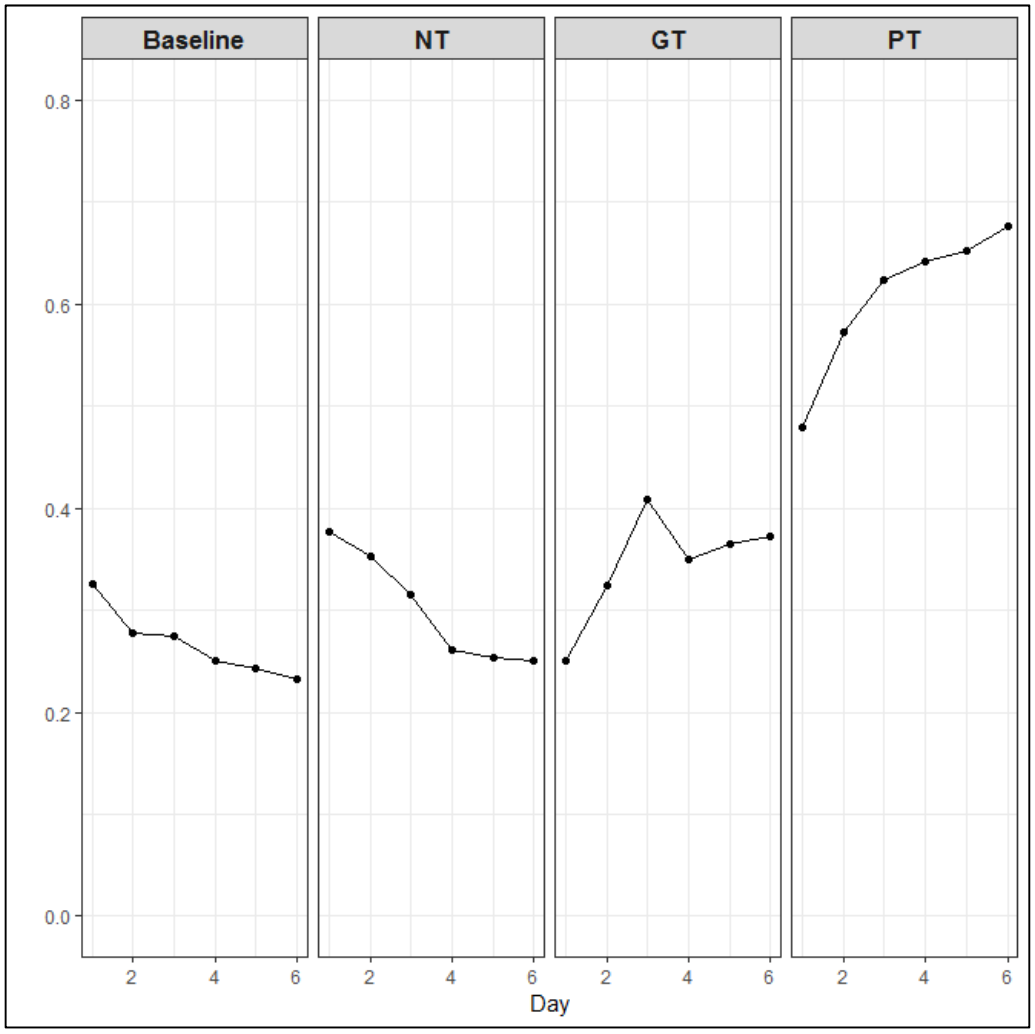
Treatment	School	Observations	Subjects	Average grade	Proportion of healthy choices
Baseline	Juan Carlos I	114	20	4.64 (1.96)	0.33 (0.23)
	Santo Domingo I	96	18	4.90 (1.73)	0.39 (0.23)
	Santo Domingo II	188	35	4.78 (2.34)	0.35 (0.29)
NT	Alfredo Cazaban	168	29	5.07 (2.05)	0.41 (0.28)
	San Isidoro I	115	20	5.49 (1.88)	0.45 (0.26)
	San Isidoro II	128	22	5.95 (1.90)	0.49 (0.26)
GT	Nuestra Señora de la Capilla	141	24	5.42 (1.73)	0.46 (0.24)
	Principe Felipe	140	25	6.09 (1.96)	0.53 (0.25)
	San Miguel	92	16	5.26 (1.74)	0.38 (0.22)
PT	Gloria Fuertes I	139	24	8.01 (1.33)	0.76 (0.24)
	Gloria Fuertes II	142	24	7.52 (1.98)	0.68 (0.28)
	Tolosa	148	25	8.09 (1.41)	0.78 (0.21)

Figure E.1. Evolution of children's average grades between days 1- 6



Notes: The figure replicates the analysis summarized in Figure 2 of the main text, but using children's average grades as a measure of healthy choices.

Figure E.2. Percentage of healthy choices over time, using a stricter definition of "healthy choice" (i.e., items that were assigned a grade of 10)



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.6. Multinomial logit model for treatment assignment, among schools that participated both in the first-stage and surprise sessions (Days 1 – 7)

Treatment	Covariate	Estimates
Nutritionist (NT)	*Constant*	1.91
		(4.32)
	Cluster-Adjusted p-values	(1.00)
	Milk and dairy products	-0.58
		(0.52)
	Cluster-Adjusted p-values	(0.92)
	Carbohydrates	-0.35
		(0.37)
	Cluster-Adjusted p-values	(0.66)
	Proteins	0.10
		(0.43)
	Cluster-Adjusted p-values	(0.28)
	Fruit and vegetables	0.14
		(0.74)
	Cluster-Adjusted p-values	(0.24)
	Fats and sugars	0.21
		(0.49)
	Cluster-Adjusted p-values	(0.21)
	Male	-0.45
		(0.72)
	Cluster-Adjusted p-values	(0.37)
	Parents hold University degree	-0.07
		(1.06)
	Cluster-Adjusted p-values	(0.32)
	Both parents are economically inactive	-1.16
		(0.84)
	Cluster-Adjusted p-values	(0.40)
	Household Income	0.76
		(0.46)
	Cluster-Adjusted p-values	(0.46)
	Involvement in children's school	0.33

	success	(0.59)
	Cluster-Adjusted p-values	(0.58)
	Average school grade	-0.28 (0.28)
	Cluster-Adjusted p-values	(0.43)
Grades (GT)	*Constant*	-4.73 (4.19)
	Cluster-Adjusted p-values	(1.00)
	Milk and dairy products	-0.98 (0.51)
	Cluster-Adjusted p-values	(0.62)
	Carbohydrates	-0.16 (0.33)
	Cluster-Adjusted p-values	(0.57)
	Proteins	0.82 (0.41)
	Cluster-Adjusted p-values	(0.20)
	Fruit and vegetables	-0.17 (0.72)
	Cluster-Adjusted p-values	(0.16)
	Fats and sugars	-0.15 (0.52)
	Cluster-Adjusted p-values	(0.14)
	Male	-0.71 (0.68)
	Cluster-Adjusted p-values	(0.75)
	Parents hold University degree	0.20 (1.12)
	Cluster-Adjusted p-values	(0.80)
	Both parents are economically inactive	1.09 (0.81)
	Cluster-Adjusted p-values	(0.42)
	Household Income	0.61 (0.44)
	Cluster-Adjusted p-values	(0.50)

	Involvement in children's school success	0.67 (0.54)
	Cluster-Adjusted p-values	(0.40)
	Average school grade	0.27 (0.27)
	Cluster-Adjusted p-values	(0.51)
Parents (PT)	*Constant*	-6.91 (4.91)
	Cluster-Adjusted p-values	(1.00)
	Milk and dairy products	-1.13 (0.56)
	Cluster-Adjusted p-values	(0.57)
	Carbohydrates	-0.40 (0.40)
	Cluster-Adjusted p-values	(0.55)
	Proteins	1.13 (0.49)
	Cluster-Adjusted p-values	(0.85)
	Fruit and vegetables	-0.61 (0.81)
	Cluster-Adjusted p-values	(0.95)
	Fats and sugars	0.05 (0.60)
	Cluster-Adjusted p-values	(0.17)
	Male	-0.84 (0.80)
	Cluster-Adjusted p-values	(0.49)
	Parents hold University degree	-0.58 (1.18)
	Cluster-Adjusted p-values	(0.49)
	Both parents are economically inactive	-0.19 (0.91)
	Cluster-Adjusted p-values	(0.38)
	Household Income	1.26 (0.50)

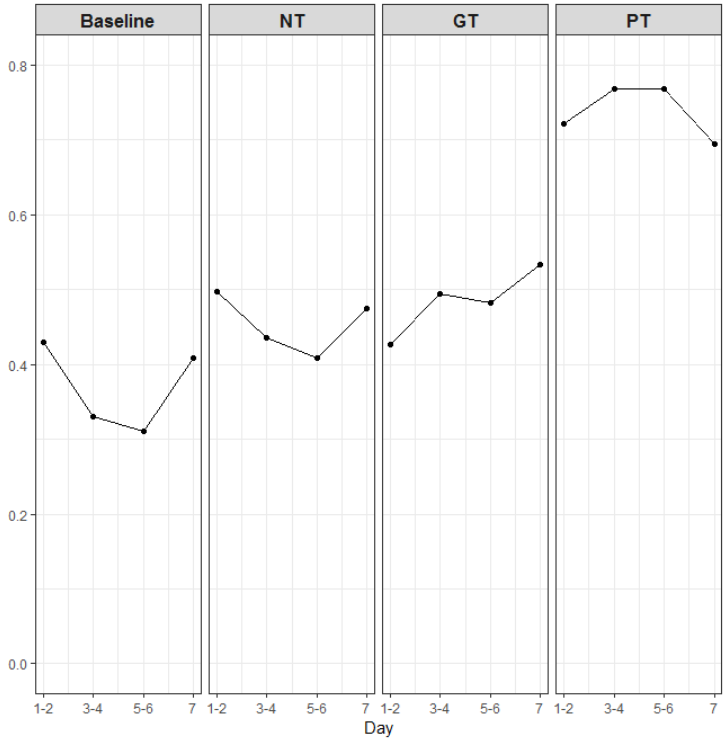
	Cluster-Adjusted p-values	(0.55)
	Involvement in children's school success	0.79 (0.61)
	Cluster-Adjusted p-values	(0.59)
	Average school grade	0.35 (0.35)
	Cluster-Adjusted p-values	(0.31)

Notes: The table reports parameter estimates from a multinomial logit model for the assignment to treatment of subjects in the 10 schools that participated in the first-stage and surprise sessions, 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,* and *Both parents are economically inactive* 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.

Table E.7. Average grades and healthy food choices in the surprise session, by school

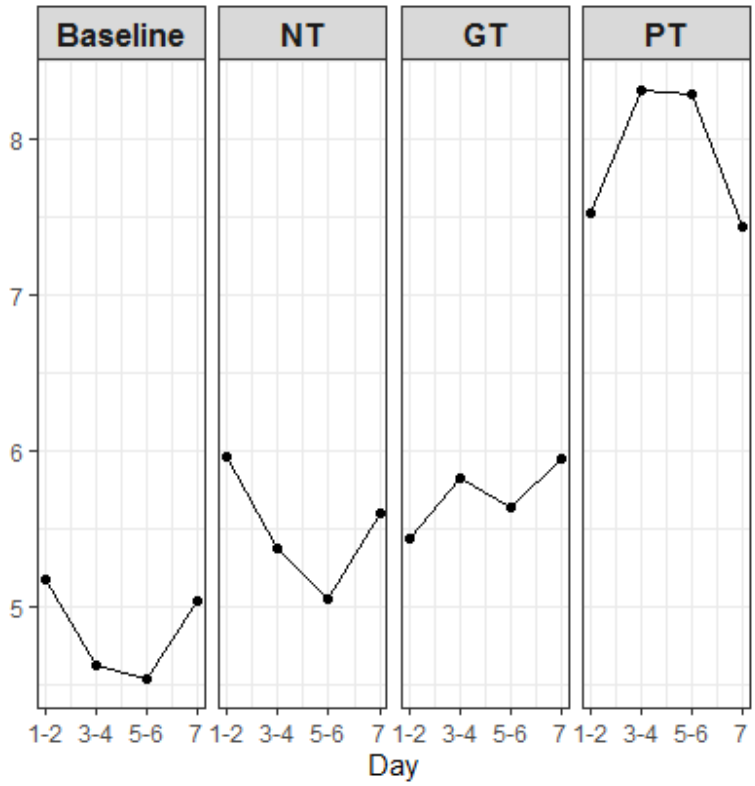
Treatment	School	Observations	Subjects	Average grade	Proportion of healthy choices
Baseline	Juan Carlos I	17	17	4.89 (1.73)	0.35 (0.25)
	Santo Domingo I	16	16	5.19 (1.70)	0.47 (0.24)
NT	Alfredo Cazaban	29	29	5.40 (1.94)	0.45 (0.28)
	San Isidoro I	18	18	5.90 (1.52)	0.51 (0.23)
	San Isidoro II	21	21	5.60 (1.75)	0.48 (0.25)
GT	Nuestra Señora de la Capilla	24	24	5.44 (1.95)	0.51 (0.28)
	Principe Felipe	25	25	6.45 (2.18)	0.58 (0.29)
	San Miguel	16	16	5.90 (1.25)	0.50 (0.23)
PT	Gloria Fuertes I	22	22	7.13 (1.87)	0.66 (0.21)
	Tolosa	23	23	7.17 (1.39)	0.73 (0.21)

Figure E.3. Percentage of healthy choices on days 1-6 and on day 7 (surprise session) including data from all the schools that took part in the first-stage sessions



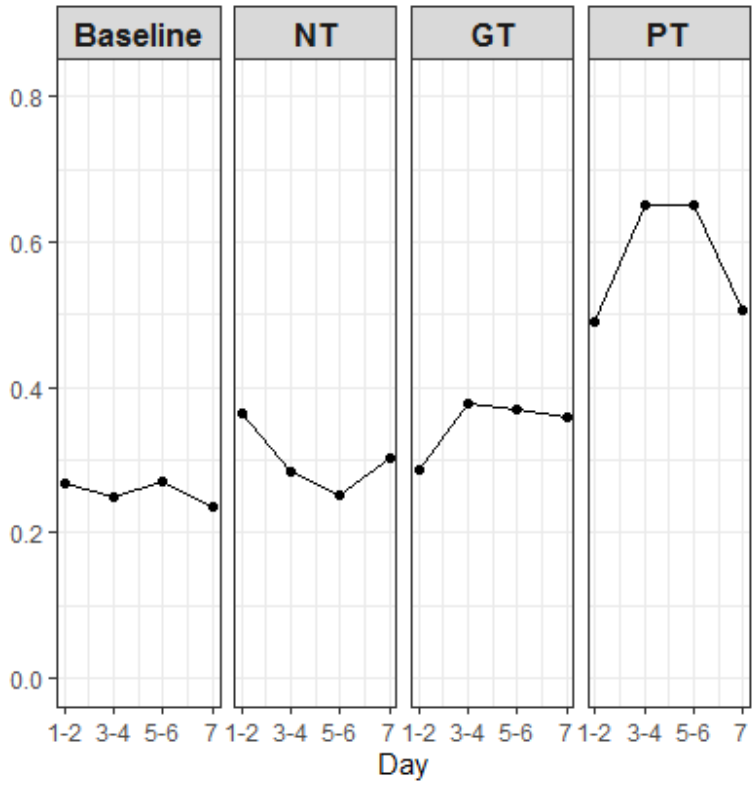
Notes: The figure replicates the analysis summarized in Figure 3 of the main text, but using data from all the schools that took part in the first-stage sessions (days 1 – 6).

Figure E.4. Children's average grades between on days 1-6 and on day 7 (surprise session)



Notes: The figure replicates the analysis summarized in Figure 3 of the main text, but but using children's average grades as a measure of healthy choices. We use only data from the schools that participated both in the first stage and in the surprise sessions.

Figure E.5. Percentage of healthy choices on days 1-6 and on day 7 (surprise session) using a stricter definition of "healthy choice" (i.e., items that were assigned a grade of 10)



Notes: The figure replicates the analysis summarized in Figure 3 of the main text, but but defining as "healthy choices" those with grades equal to 10 only. We use only data from the schools that participated both in the first stage and in the surprise sessions.

Table E.8. Multi-level probit analysis on the probability of choosing healthy food items

	(1)	(2)	(3)
Constant	-1.34*** (0.09)	-1.44*** (0.13)	-1.48*** (0.18)
NT	0.27** (0.11)	0.35*** (0.14)	0.28* (0.16)
GT	0.20* (0.12)	0.34** (0.16)	0.39** (0.18)
PT	1.39*** (0.11)	1.39*** (0.13)	1.34*** (0.16)
Male		0.10 (0.08)	0.15* (0.09)
Average school grade		0.30*** (0.09)	0.21** (0.11)
Personal Impulsivity		-0.18* (0.09)	-0.21* (0.10)
School Impulsivity		-0.13 (0.08)	-0.07 (0.09)
Parents hold University degree			0.05 (0.11)
Household Income			0.01 (0.11)
#Observations	1,611	1,347	1,129
Deviance Information Criterion	1,463.98	1,280.86	1,089.39

Notes: The table reports parameter estimates from multi-level probit 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.9. Regression analysis of subjects' average grades

	(1)	(2)	(3)
Constant	4.78*** (0.10) [0.00]	4.69*** (0.47) [0.00]	4.97*** (0.57) [0.00]
NT	0.68*** (0.14) [0.02]	0.79*** (0.16) [0.04]	0.87*** (0.18) [0.04]
GT	0.86*** (0.14) [0.00]	1.03*** (0.18) [0.03]	1.12*** (0.21) [0.02]
PT	3.09*** (0.13) [0.00]	3.03*** (0.15) [0.01]	3.17*** (0.18) [0.00]
Male		0.13 (0.10) [0.16]	0.15 (0.11) [0.23]
Average school grade		0.16*** (0.04) [0.05]	0.14** (0.05) [0.07]
Personal Impulsivity		-0.32*** (0.08) [0.05]	-0.38*** (0.09) [0.04]
School Impulsivity		-0.18* (0.09) [0.11]	-0.15 (0.11) [0.31]
Parents hold University degree			-0.19 (0.14) [0.59]
Household Income			-0.04 (0.15) [0.34]
#Observations	1,611	1,347	1,129
R^2	0.28	0.33	0.33

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.

Table E.10. Multi-level regression analysis of subjects' average grades

	(1)	(2)	(3)
Constant	4.65*** (0.33)	5.86*** (0.56)	4.92*** (0.59)
GT	1.04** (0.43)	1.22** (0.58)	1.59** (0.58)
NT	0.82** (0.41)	1.02** (0.47)	0.94* (0.52)
PT	3.20*** (0.40)	3.14*** (0.41)	3.12*** (0.51)
Male		0.28 (0.22)	0.47 (0.26)
Average school grade		0.16*** (0.06)	0.12 (0.08)
Personal Impulsivity		-0.64 (0.47)	-1.13** (0.56)
School Impulsivity		-0.29* (0.15)	-0.22 (0.17)
Parents hold University degree			-0.14 (0.23)
Household Income			0.49*** (0.14)
# Observations	1,611	1,347	1,129
Deviance Information Criterion	5,129.64	5,025.14	3,324.99

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.11. Probit model for the daily differences in the probability of making healthy food choices across treatments

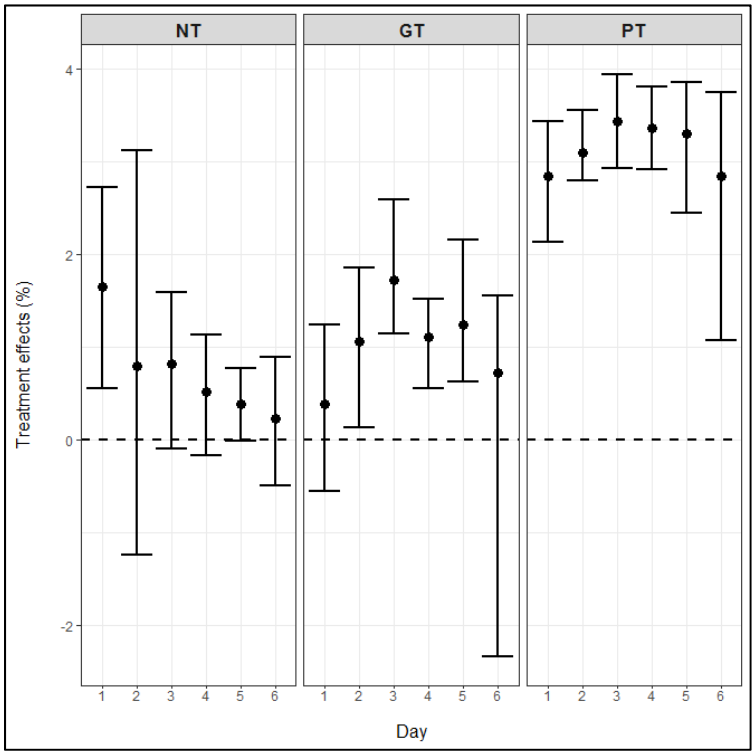
Covariates	Parameter estimates	Score bootstrap
Constant	-0.12 (0.02)	z=-1.46, p=0.07
PT x Day 1	1.72 (0.11)	z=2.77, p=0.01
PT x Day 2	1.85 (0.23)	z=2.75, p=0.00
PT x Day 3	2.04 (0.37)	z=2.76, p=0.03
PT x Day 4	2.32 (0.34)	z=2.77, p=0.00
PT x Day 5	2.32 (0.35)	z=2.76, p=0.00
PT x Day 6	1.71 (0.43)	z=2.67, p=0.00
GT x Day 1	0.16 (0.11)	z=0.98, p=0.53
GT x Day 2	0.82 (0.17)	z=1.58, p=0.00
GT x Day 3	1.00 (0.44)	z=1.46, p=0.00
GT x Day 4	0.53 (0.15)	z=1.48, p=0.00
GT x Day 5	1.02 (0.16)	z=1.66, p=0.00
GT x Day 6	0.31 (0.15)	z=1.19, p=0.13
NT x Day 1	0.90 (0.15)	z=2.19, p=0.00
NT x Day 2	0.47 (0.24)	z=1.51, p=0.14
NT x Day 3	0.51 (0.14)	z=1.67, p=0.13
NT x Day 4	0.38 (0.18)	z=1.18, p=0.27
NT x Day 5	0.38 (0.10)	z=0.49, p=0.55
NT x Day 6	0.12 (0.17)	z=0.78, p=0.46

Notes: The table summarizes the results from the probit models used as a basis to produce Figure 4 in the main text. *Parameter estimates* Column (middle column) reports the parameter estimates; cluster-robust standard errors clustered by school in parentheses. *Score bootstrap* Column (right column) reports t-statistic and p-values (in parentheses) for two-sided Wald tests computed according to Kline and Santos' (2012) score bootstrap procedure accounting for small number of clusters (schools).

Covariates	Parameter estimates	Score bootstrap
Constant	4.72 (0.05)	
PT x Day 1	2.84 (0.26)	t=10.90, p=0.01
PT x Day 2	3.11 (0.12)	t=24.99, p=0.00
PT x Day 3	3.43 (0.21)	t=16.24, p=0.00
PT x Day 4	3.37 (0.16)	t=20.78, p=0.00
PT x Day 5	3.30 (0.30)	t=11.05, p=0.00
PT x Day 6	2.85 (0.60)	t=4.71, p=0.00
GT x Day 1	0.39 (0.08)	t=4.98, p=0.16
GT x Day 2	1.07 (0.49)	t=2.20, p=0.00
GT x Day 3	1.73 (0.34)	t=5.11, p=0.00
GT x Day 4	1.11 (0.29)	t=3.85, p=0.05
GT x Day 5	1.25 (0.35)	t=3.61, p=0.02
GT x Day 6	0.73 (0.40)	t=1.83, p=0.30
NT x Day 1	1.65 (0.27)	t=6.11, p=0.03
NT x Day 2	0.81 (0.53)	t=1.51, p=0.30
NT x Day 3	0.82 (0.21)	t=4.54, p=0.13
NT x Day 4	0.52 (0.16)	t=4.54, p=0.15
NT x Day 5	0.39 (0.10)	t=2.27, p=0.39
NT x Day 6	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 E.6 below. *Parameter estimates* Column (middle column) reports the OLS estimates; cluster-robust standard errors clustered by school in parentheses. *Score bootstrap* Column (right column) 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).

Figure E.6. Daily differences in subjects' grades between each treatment (*NT*, *GT*, and *PT*) and the *Baseline* during the intervention period



Notes: The figure replicates the analysis summarized in Table 3 of the main text, but using subjects' grades as the outcome variable. Solid circles represent point estimates; vertical lines give the 95% confidence intervals, obtained by inverting the wild bootstrap-*t* tests.

Table E.12. Percentage Individual Consistency, Days 2-6 and Comparisons (8 items and 75% threshold)

Treatment	[1] Days 2-6	[2] Day 1 vs. Days 2-6	[3] Day 7 vs. Days 2-6
Baseline	71.43	28.57	61.90
NT	62.82	51.28	73.07
GT	66.17	29.41	57.35
PT	87.75	53.06	79.59

Notes: The table replicates the analysis reported in Table 4. In this case, we consider 8 items instead of 6 to establish the pre-determined pool. We keep the consistency threshold at 75%.

Table E.13. Percentage Individual Consistency, Days 2-6 and Comparisons (8 items and 100% threshold)

Treatment	[1] Days 2-6	[2] Day 1 vs. Days 2-6	[3] Day 7 vs. Days 2-6
Baseline	42.86	14.28	42.86
NT	34.61	16.67	43.58
GT	38.23	4.41	19.11
PT	71.43	20.41	40.82

Notes: The table replicates the analysis reported in Table 4. In this case, we consider 8 items instead of 6 to establish the pre-determined pool. We increase the consistency threshold to 100%

Table E.14. Percentage Individual Consistency, Days 2-6 and Comparisons (6 items and 100% threshold)

Treatment	[1] Days 2-6	[2] Day 1 vs. Days 2-6	[3] Day 7 vs. Days 2-6
Baseline	19.05	4.76	28.57
NT	8.97	12.82	32.05
GT	8.82	2.94	8.82
PT	20.41	10.20	24.45

Notes: The table replicates the analysis reported in Table 4. In this case, we also consider 6 items to establish the pre-determined pool. We increase the consistency threshold to 100%