Helping consumers with a front-of-pack label: numbers or colours? Experimental comparison between Guideline Daily Amount and Traffic Light in a diet building exercice

Paolo Crosetto, Laurent Muller, Bernard Ruffieux

Laboratoire d'Economie Appliquée de Grenoble

HELPING CONSUMERS WITH A FRONT-OF-PACK LABEL: NUMBERS OR COLOURS?

EXPERIMENTAL COMPARISON BETWEEN GUIDELINE DAILY AMOUNT AND TRAFFIC LIGHT IN A DIET-BUILDING EXERCICE

CROSETTO Paolo ; MULLER Laurent ; RUFFIEUX Bernard

- June 30, 2015 -

JEL CODES D12 ; D18 ; C91 ; C93

PSYCINFO CLASSIFICATION: 3920

Working Paper GAEL ; 2015-10

Institut National de la Recherche Agronomique - Université Pierre Mendès France
BP 47 - 38040 GRENOBLE Cedex 9 – France - Tél. : 33 (0) 4 76 82 54 40
Fax : 33 (0) 4 76 82 54 55 - Courriel : vertier@grenoble.Inra.fr - http://www.grenoble.inra.fr

Helping consumers with a front-of-pack label: numbers or colours?

Experimental comparison between Guideline Daily Amount and Traffic Light in a diet-building exercise. ☆

Paolo Crosetto[a], Laurent Muller[a], Bernard Ruffieux[b]

[a] *INRA and Univ. Grenoble Alpes, UMR 1215 GAEL, F-38000 Grenoble, France*
[b] *INP and Univ. Grenoble Alpes, UMR 1215 GAEL, F-38000 Grenoble, France*

Abstract

This paper contributes to the debate on front-of-pack nutritional labels. Because of their dissimilar formats, Guideline Daily Amount (GDA) and Traffic Light (TL) may trigger different responses among consumers. While GDA is comprehensive and cognitively demanding, information is coarser and more salient in TL. We implement an incentivized laboratory experiment to assess the relative performance of GDA and TL labelling schemes in assisting consumers to build a healthy daily diet. Participants must compose a daily diet, choosing from a finite set of products, and are paid a fixed cash amount only if the diet satisfies pre-determined nutritional goals. Goals correspond to the guideline daily amount values for different nutritional attributes, whose number varies from 1 (kcal) to 7 (kcal, fat, sugar, salt, fibre, vitamin C and calcium). Three different labels, GDA, TL and a combined GDATL are provided. Results show that GDA performs better than TL when subjects do not face time constraints. When time is limited however, TL and GDA have identical efficacy with 4 nutritional goals, and TL even outperforms GDA with 7 nutritional goals.

JEL Classifications: D12, D18, C91, C93

PsycINFO Classifications: 3920

Keywords: Nutritional labels, Food choice, Experimental Economics, Guideline Daily Amount, Traffic Lights

☆We wish to thank Marie Cronfalt-Godet for excellent research support, Jean-Loup Dupuis for invaluable logistic and IT help in running the experimental sessions, and Mariane Damois for the recruitment campaign. We gratefully acknowledge the financial support of the INRA Metaprogramme DID'IT – project FOODPOL. The paper benefited from comments and insights provided by participants to seminars and conferences in Marseille, York, Lausanne, Grenoble, Naples, Nottingham and Paris. All remaining errors are ours.

Contact: `paolo.crosetto@grenoble.inra.fr` (Paolo Crosetto), `laurent.muller@grenoble.inra.fr` (Laurent Muller), `bernard.ruffieux@grenoble-inp.fr` (Bernard Ruffieux)

Introduction

The growing awareness about the ill effects of unhealthy diets has led to the development of normative front-of-pack labels, to be added to the existing descriptive back-of-pack nutrition panels. Front-of-pack labels aim to help consumers make healthier choices. While several dozens of label formats are already in use (Drichoutis et al., 2011), the debate over their relative efficacy mainly focuses on two widespread labelling schemes: *Guideline Daily Amounts* (GDA) and *Traffic Lights* (TL). Both labels give information about the approximate amount of calories, fat, saturated fat, total sugars and salt[1]. While GDA expresses the information as percentage of a recommended daily value, TL display color-coded information using a 3-color scale (green, amber, red) derived from road lights. GDA and TL can be combined; we will refer such label in the following as GDATL. Originally called Daily Guideline Intakes by the British Food Standards Agency (1996), GDA has been recently replaced by *Reference Intake*[2] in the UK. The GDA system has been adopted by the Australian food and beverage industry in 2006 (*Daily Intake Guide*), by the European Union in 2009 as an industry standard and has been introduced in the US following Michelle Obama's initiative in 2012 (*Facts Up Front*). Although widespread, the GDA system may be difficult to understand for many and does not lend to quick comparisons. TL have been proposed by the British Food Standard Agency in order to make the information easily and rapidly understood. While TL are supported by the British Medical Association and welcomed by consumers, the food industry worries that foods marked red would be shunned and qualifies this system as too simplistic, misleading, patronising and unscientific.

We focus here on the mere efficiency of the label, abstracting away from exposure, perception, taste, understanding, preferences, and actual use (for recent reviews about existing research on several aspects of the relationship between consumers and labels see Grunert and Wills, 2007; Vyth et al., 2012). We implement an incentivized laboratory experiment to assess the relative performance of GDA, TL or GDATL in helping consumers build a healthy daily diet. We give the subjects clear and unambiguous nutritional goals in order to test, in a controlled environment, which label, if any, leads to the most efficient behaviour.

Many studies have fueled the debate about what format between GDA and TL should be favoured (see Kelly et al. (2009); Moeser et al. (2010); Grunert et al. (2010) for experimental studies and Grunert and Wills (2007); Hawley et al. (2013) for reviews of the literature). While these studies are highly valuable to assess how consumers perceive, understand and use GDA and TL, they suffer from two shortcomings when it comes to the question of the labels' performance in assisting the consumer to make healthy food choices.

First, evaluations of performance are based on products rather than diet. The proclaimed objective of GDA and TL is to get consumers' overall balance right. Most experimental studies, though, implement simplified choice environments, in which consumers have to rank two or maximum three products ac-

[1] GDA and TL also both provide the absolute amount per serving of these categories.

[2] Although the principles behind both are the same, the major difference is that GDA existed for men, women and children; there is only one set of *Reference Intakes* for an average adult.

cording to their perceived healthiness. These exercises are informative, since the aggregation of healthy products results in a healthy diet (the converse is not necessarily true, though). Nonetheless, constructing a healthy diet is a different task than facing binary choices: on a daily basis, consumers must select dozens of food items so that the sum of their nutrients meet predetermined targets. In fact, consumers are asked to perform a sort of algebraic exercise.

Second, a thorough survey of the literature reveals that the *question asked* determines the relative performance of the labelling schemes. Questions seem to favour TL when they are ordinal and involve small choice sets. For instance, when the subjects are asked to rank the products' *relative* healthiness, or to classify them into three-level scales as healthy/medium/unhealthy, TL wins (see, for instance, Kelly et al., 2009; Borgmeier and Westenhoefer, 2009). On the other hand, questions seem to favour GDA when they are cardinal and involve large choice sets. When the subjects are asked to provide *absolute* assessments, i.e. to evaluate how much of a nutrient is present in each product, GDA wins (see Synovate, 2005). This is not surprising: as already noted by Grunert and Wills (2007) labels perform best when the question asked is the one they have been designed to answer.

Our experimental setting directly relates to the labels' primary objective: Subjects were asked to build daily menus by choosing from a large set of products and within a predetermined meal structure. Subjects were paid if and only if the chosen menu satisfied a known and well-defined set of nutritional criteria. To guide subjects in their choices, GDA, TL or GDATL were provided. We also varied, within subjects, the number of nutritional goals. The participants faced easy, 1-dimensional tasks, in which the daily menus had to satisfy only an energy constraint; medium, 4-dimensional tasks, in which goals included energy but also limits on the amount of bad nutrients (saturated fat, sugar and salt); and difficult 7-dimensional tasks, in which on top of the above participants had also to maximise the amount of good nutrients, namely vitamin C, calcium and fibre.

The experiment was designed so that GDA and TL are only assessed on their efficacy to help consumers build a healthy daily diet. We abstract away from consideration about the salience and use of the labels in a real shopping environment. In order to observe decisions that are independent of personal taste and preferences, subjects were asked to act as hired nutritionists of a refectory that catered to all sorts of people for the whole day and earned money for each menu created that satisfied the given constraints. The experimental task was also a simplification of the actual consumer problem. Supermarket shelves comprise thousands of products that are competing for the consumers' limited attention. Furthermore, consumers must keep a running memory of what they have bought, in order to fully take into account the overall nutritional balance of their shopping. In the laboratory, subjects are not distracted and are fully dedicated to the assigned task. Besides, we presented all information and all food choices for the whole day in one compact and intuitive screen, in order to limit the effects of working memory constraints (see a vast literature on the subject started by Miller, 1956). These design choices limit the external validity of our results, but allow for a tightly controlled and neutral testing environment.

What should be expected from this experiment? The experimental setting clearly favours GDA. Thanks to numerical information in the form of percentages, GDA is best suited to compute daily amounts. With a three-color scheme, TL gives coarser information, and is hence strictly inferior from a computing point of view since it provides three intervals rather than continuous values. Furthermore,

the relative appeal of TL compared to GDA fades away in our laboratory context: subjects' attention is captured *per se* and monetary incentives are conducive to hard reasoning. Should GDA fail in such an environment, we could hence conclude that its application in the real world would be fragile at best. In order to be effective, GDA requires high computational skills, especially when the nutritional goals are multiple. While GDA would fit well to *homo oeconomicus*, TL may be more beneficial to less careful and unskilled decisions makers: TL arguably requires lower cognitive skills as it appeals to the intuitive side of human decision making. It has also been observed that too much information can lead to bad choices (Greifeneder et al., 2010; Malhotra, 1982), especially if the information is multidimensional. TL labels could hence generate better results when used by cognitively constrained consumers.

We ran two different experiments with the same overall structure. In a first experiment, we give subjects unlimited time and paper and pencil to perform calculations. We recruited two different samples of subjects: highly skilled and math-oriented engineering students and a representative sample of the general population. We proposed 15 different daily diet tasks to both sets of subjects, spanning 1, 4 and 7 dimensions, and measured their performance with respect to the given task and the time spent.

Results of both the student and general population samples show that GDA performs better than both TL and the combined GDATL, on both 4- and 7-dimensional tasks. Time spent on each task is high (3 minutes on average). TL leads to slightly faster decisions than GDA, while GDATL leads to slower decisions. Both samples performed extensive computations, the main difference being that engineering students are, not surprisingly, better at constrained optimization than subjects taken from the population at large. Experiment 1 shows that GDA performs hands down better than TL. This is more the case with highly-skilled, rational engineer students than for the general population.

In a second experiment we test the robustness of these findings, and make two steps towards external validity: we imposed a time limit of 2 minutes per task and did not give the subjects paper and pencil or any other tool to perform computations. We recruited subjects from the population at large. We also added two sets of controls: two image-only tasks in which no nutritional information was given in order to examine the distance between the recommended healthy diet with diets based on subjects' (i) preference and (ii) beliefs, and two pure mathematical tasks in order to assess technical skills in simple arithmetic computations.

In Experiment 2, with 4 nutritional goals GDATL performs better than both TL and GDA, who perform equally. With 7 nutritional goals, both GDATL and TL outperform GDA. The effect is robust to controlling for subject's preferences and performance in the mathematical and preference-based tasks. Experiment 2 shows that introducing some minor constraints to the subjects, in a setting still largely biased towards GDA, suffices to weaken the performance of GDA to the point that GDATL performs uniformly better, and TL performs equally or better depending on task complexity.

The combined results of our two experiments seem to suggest that while GDA is the right tool in the hands of unconstrained, highly-trained individuals with time, focus, and incentives, it falls short to the theoretically inferior TL as soon as some limited constraints are imposed on the subjects. While we cannot claim that our results generalize outside of the lab, we see them as a strong hint to the fact that, if GDA fails in a setting strongly biased in its favour, then it might only do worse in the case of real purchases, in which consumers are severely bounded in time, attention, focus, and budget.

Experiment 1 - Benchmark

Methods

Experimental design and treatments

The aim of the first experiment was to build a controlled environment to test the relative efficacy of three different labelling schemes (GDA, TL, GDATL) combined with three different levels of complexity (1,4, and 7 nutritional goals).

We told subjects to act as hired nutritionist of a refectory, that catered for all sorts of people for the whole day. Subjects had to compose daily diets that satisfied a set of pre-determined nutritional criteria. They had to choose among the finite set of food items that the refectory had in store. A daily diet was made up of twelve food items laid out in a fixed French-style meal structure: breakfast, three-course lunch, afternoon snack, three-course dinner. For each component, subjects had to pick one food item out of four available alternatives. A treatment-specific nutritional label was provided for each food item. Subjects knew from the outset that they had to repeat this task 15 times, with different products on each repetition, and that they would be paid €1.50 for each successfully completed task (i.e., if the composed daily diet satisfied all nutritional goals).

The number of nutritional goals varied from 1 to 7. In one-dimensional tasks (1D) subjects had to keep the caloric content within bounds. In four-dimensional tasks (4D), bad nutrients had to be kept below a pre-determined limit. In seven-dimensional tasks (7D), good nutrients had to be provided in sufficient amount. All constraints were set and communicated to the subjects as percentages of a daily recommended amount; their details are given in Table 1.

Dimensions	**Nutrient**	**Threshold (as % of recommended daily amount)**
1D	Kcal	$90 \leq \sum \text{Kcal} \leq 110$
4D	Sugar	$\sum \text{Sugar} \leq 100$
	Saturated fat	$\sum \text{Fat} \leq 100$
	Salt	$\sum \text{Salt} \leq 100$
7D	Fiber	$\sum \text{Fiber} \geq 100$
	Vitamin C	$\sum \text{Vitamin C} \geq 100$
	Calcium	$\sum \text{Calcium} \geq 100$

Table 1: Constraints given to subjects for daily diets, by dimension

Each task was displayed on a single screen. One screen contains 4 food items for each of the 12 components of the daily diet. We used 5 different screens that included different products. Each screen was repeated three times, once in 1D, once in 4D and once in 7D. The position of the food items was scrambled across repetitions to reduce the likelihood that subject recognized to be in the same screen again. An example of nutrition screen with 4D and combined GDATL label is given in Figure 1.

Subjects had unlimited time on each screen, and were provided with paper and pencil, to allow them to take notes and/or perform computations if they wished to do so.

Figure 1: Screenshot of a 4D, GDATL decision screen

We chose a mixed between-within design. Labels varied between subjects, while the number of nutritional goals varied within-subjects: Each subject faced only one label format, over three degrees of complexity.

Procedures and materials

We gathered full nutritional data for 346 widely consumed general food items. The nutritional data came from the Su-Vi-Max project (INSERM, 2006)[3] complemented with data from the Darmon et al. (2009) study. We coded each food item as belonging to a food family (meat, fish, fruits, vegetables...), and assigned it to one or more daily meals (breakfast, lunch, afternoon snack, dinner) following French eating habits. Each food item is represented by its ready-to-eat picture, centered on a neutral white background.[4]

Recommended daily amounts were generated from the raw nutritional data following European Union (2011) official tables, with reference to an average adult.[5] The Traffic Lights color thresholds were in turn computed from the recommended daily amount, applying a simplified version of those in force in the UK (FSA, 2013). For bad nutrients, such as salt, sugar and saturated fat, the traffic light would be green for contents lower or equal to 5% of the recommended daily amount, red for a content of more than 25%, and amber otherwise. For good nutrients such as vitamins, fiber and calcium we inverted the thresholds, assigning green to food containing more than 25% of the recommended daily amount, and red if less than 5%. This simplification was needed to give a simple and clear message

[3]The Su-Vi-Max database gives the nutritional content (39 components) of 923 food items commonly eaten by French adults.

[4]We decided to display ready-to-eat rather than packaged products to avoid trademarks and the nutritional beliefs, habits, information on prices and social desirability, that go along with them.

[5]The full conversion table is available in Appendix AppendixA.

to our participants, with no substantial differences with reality[6]. The used threshold are summarised in Table 2.

	% of recommended daily amount		
	5%		25%
Salt	Green	Amber	Red
Sugar	Green	Amber	Red
Fat	Green	Amber	Red
Vitamin C	Red	Amber	Green
Fiber	Red	Amber	Green
Calcium	Red	Amber	Green

Table 2: TL threshold employed in the experiment for all nutrients

We used our product database to build possible daily diets, i.e., lists of food items that make up for a full day of consumption. The daily diets were built with reference to traditional French eating habits, and were composed of a light continental breakfast, lunch (starter, main course, dessert), afternoon snack, and dinner (starter, main course, side and dessert). To balance the diet with commonly used items such as bread, oil or butter, we added to the diet a *daily base* composed of 120 grams of white bread, 20 grams of oil and 10 of butter. A daily diet was hence overall composed of 11 food items and a common base. Examples of which food items were assigned to each meal are given in Table 3

Daily base	-	120g bread, 10g butter, 20g oil
Breakfast	*Drink*	Tea, coffee, milk, hot chocolate, juice...
	Main course	Bread, sweets, *viennoiseries*...
	Fruit	Fruit, jam...
Lunch	*Starter*	Light dishes, ham, paté...
	Main course	Sandwich, pizza, pasta...
	Dessert	Fruit, sweets...
Afternoon snack	-	Sweets
Dinner	*Starter*	Light dishes, ham, paté...
	Main course	Meat or fish
	Side	Vegetables, rice, potatoes...
	Dessert	Fruit, sweets...

Table 3: The chosen structure of a daily diet, with example food items

We randomly generated several thousand daily diets, and then checked them against nutritional criteria, with the aim of singling out both healthy and unhealthy diets. We considered a daily diet healthy if

[6]The actual thresholds in use in the United Kingdom (see Appendix AppendixA) slightly vary by nutrient but are by and large around the chosen 5 and 25% thresholds.

(a) Kcal-only

(b) GDA

(c) TL

(d) GDATL

Figure 2: Toast & Jam: different formats for the nutritional information. 7-d not shown.

it simultaneously satisfied all 7 nutritional goals. That is, a daily diet was deemed healthy if it contained, over the 12 components, between 90% and 110% of the daily recommended amount of calories, weakly less than 100% of the daily recommended amount of salt, sugar and saturated fat, and weakly more than 100% of the daily recommended amount of vitamin C, calcium and fiber. We considered diets Unhealthy if they failed all seven criteria. After checking the diets for abnormal food items (i.e., items that would seem out of place for French eating habits), we selected 12 healthy and 12 unhealthy diets. Participants faced decision screens made of four daily diets, 2 healthy and 2 unhealthy. This generated screens made up of four columns (the four daily diets) and 12 rows (the components of a daily diet).

The information content of the screens varied across treatments. In the GDA treatment, all information was displayed as a % of the recommended daily amount of the relevant nutrient. In the TL treatment, the information appeared as color-code only. In the combined GDATL label, numerical information was shown alongside with the TL color-coding. The energy content was always expressed as % of the recommended daily amount and never as colour codes (even with TL and GDATL: As people need to take in a minimum of calories to survive, meeting the caloric constraint would not be guaranteed, unlike for bad and good nutrients, by collecting only green-patched items (see Figure 2).

We generated all the possible screens from our 24 starting daily diets, and then selected the best among them as our final decision screens. We followed two rules in this selection process. First, screens had to avoid giving the subjects "odd" choices, given the average French eating habits. An example of an "odd choice" would be to have 4 too close substitutes in the same row (4 types of bread, for example), or to have a clearly dominant option (*croissants* faced with three obscure substitutes). We checked the screens manually, and in all "odd" situations we dropped the screen from consideration. Second, we checked via numeric simulation if the screens were "too easy" or "too hard" given the 1, 4 or 7 constraints. To do so, we simulated random play on the screens, and we checked the probability that a random player would satisfy the 1, 4, or 7 constraints, screen by screen. We kept only those screens that had a middle range of probability of success.

The end result of all these procedures were five final decision screens, for a total of 20 (5×4) daily diets and 240 (20×12) products. Products were not allowed to appear twice on the same screen, but could be repeated across screens. The scores of the random-play simulations on these 5 screens are given in Table 4. Passing a single constraint (Kcal) is rather easy for a random player, but random agents

have just about a chance in 30 of simultaneously satisfying 4 constraints, and one in 250 for 7 constraints (last two columns of Table 4).

	Probability (%) of satisfying the constraint when playing randomly								
	Kcal	*Fat*	*Sugar*	*Salt*	*Fibre*	*Vitamin C*	*Calcium*	*4D*	*7D*
Screen 1	29.68	48.18	68.71	72.98	51.70	58.57	47.91	4.61	0.57
Screen 2	32.78	57.84	70.27	58.76	51.00	48.29	40.03	4.78	0.49
Screen 3	30.64	53.75	56.36	67.23	52.50	64.91	45.11	2.54	0.49
Screen 4	31.87	49.03	60.71	67.56	42.14	61.39	43.69	2.98	0.30
Screen 5	29.98	25.85	65.66	49.40	35.98	58.36	24.94	0.86	0.08
Average	30.99	46.93	64.34	63.19	46.66	58.30	40.34	3.15	0.39

Table 4: Performance of random players on the final screens

Participants

We recruited participants from two distinct subject pools: engineering students from the National Polytechnic Institute of Grenoble (INP), and individuals from the general population. The INP students belong to the top-tier engineer students in France and are at ease with mathematics. Students were recruited via ads on campus and communication at economics courses at the INP. Subjects from the general population were recruited with ads on local newspapers as well as from an existing database of potential subjects in and around Grenoble.

A total of 86 subjects took part to 6 experimental sessions: 47 INP students over 3 sessions and 39 participants from the general population over 3 further sessions.

The experimental sessions were ran in the GAEL laboratory of experimental economics, located within the INP engineering school in the centre of Grenoble, France. The experimental software was programmed in PHP, and access to the source code is available upon request. The English translation of the original French experimental instructions is available upon request.

Incentives

Participants received a show-up fee of 10 €. On top of this amount, participants could earn additional money by correctly performing the tasks. Participants were faced with 15 choice screens: five each for 1D, 4D and 7D tasks. For each task that they completed successfully, subjects earned 1.5€. Theoretical final earnings could range from 10 to 32.5 euro, for an experiment that lasted on average one hour and a half. Participants were given no feedback about the success or failure of their tasks. The screens followed one another, and participants knew their total earnings only at the end of the experiment.

Measures

We exposed the subject to the main task described above and to a questionnaire. Our key experimental measures in the main tasks were *success rate*, *distance*, and *time*.

Success rate . We measure if the subject successfully created a daily diet satisfying the given constraints or not. This is a binary measure, taking values 0 or 1 for each repetition of the task for each subject.

Distance . We computed the absolute distance, in percentage points, of the subject from each target. Any result within the target was noted as a distance of zero; farther away results were translated into positive distances. For instance, in 1D tasks, this means the distance from the 90% - 110% interval set as target, and reaching both 80% and 120% means a distance of 10. Euclidean distances were used to aggregate multi-dimensional tasks.

Time . We measured the time spent on each task by each subject, in milliseconds.

Questionnaire . The final questionnaire included questions about age, gender, income, education, and a set of questions about eating habits, including snacking, the amount of money spent weekly on food, the time allocated to cooking for each meal of the day.

Results

The average results of our measures by subject pool, label and dimensions are displayed in Table 5. The formal statistical test of hypotheses, carried out using the Mann-Whitney non-parametric test, is reported in detail for all pairs of treatments in Appendix AppendixB, in Table B.11 for INP students and in Table B.12 for the general population. Differences between students and general population are significant across the board (not reported).

		Students			General Population		
Dimensions	**Treatment**	**Success rate**	**Distance**	**Time (sec)**	**Success rate**	**Distance**	**Time (sec)**
1	GDA	0.95	0.49	82.77	0.71	3.49	118.85
	GDATL	0.94	0.70	85.19	0.74	3.09	172.46
	TL	0.93	0.56	89.77	0.78	3.27	120.55
4	GDA	0.91	1.08	180.76	0.56	7.60	223.54
	GDATL	0.84	1.28	213.91	0.48	7.52	243.22
	TL	0.55	13.21	168.74	0.35	19.09	167.55
7	GDA	0.86	3.23	251.81	0.41	13.06	310.47
	GDATL	0.78	3.01	346.44	0.29	19.54	319.77
	TL	0.36	22.28	225.10	0.12	32.52	190.78

Table 5: Experiment 1 main variables - treatment averages

Results are also reported in graphical form. The treatment means of the performance in the task is reported in Figure 3, top. The figure reports on the vertical axis the percentage of tasks correctly carried out; the error bars represent the standard error of the mean. The dotted lines in the plot represent the average expected performance of a random agent.

Students are extremely good at the task. Task complexity affects the results, but in a slight way, at least for GDA tasks. Not only students diverge significantly from random play, but they get very close to submitting 100% of correct decisions in 1D tasks. The performance on 1D task is similar across

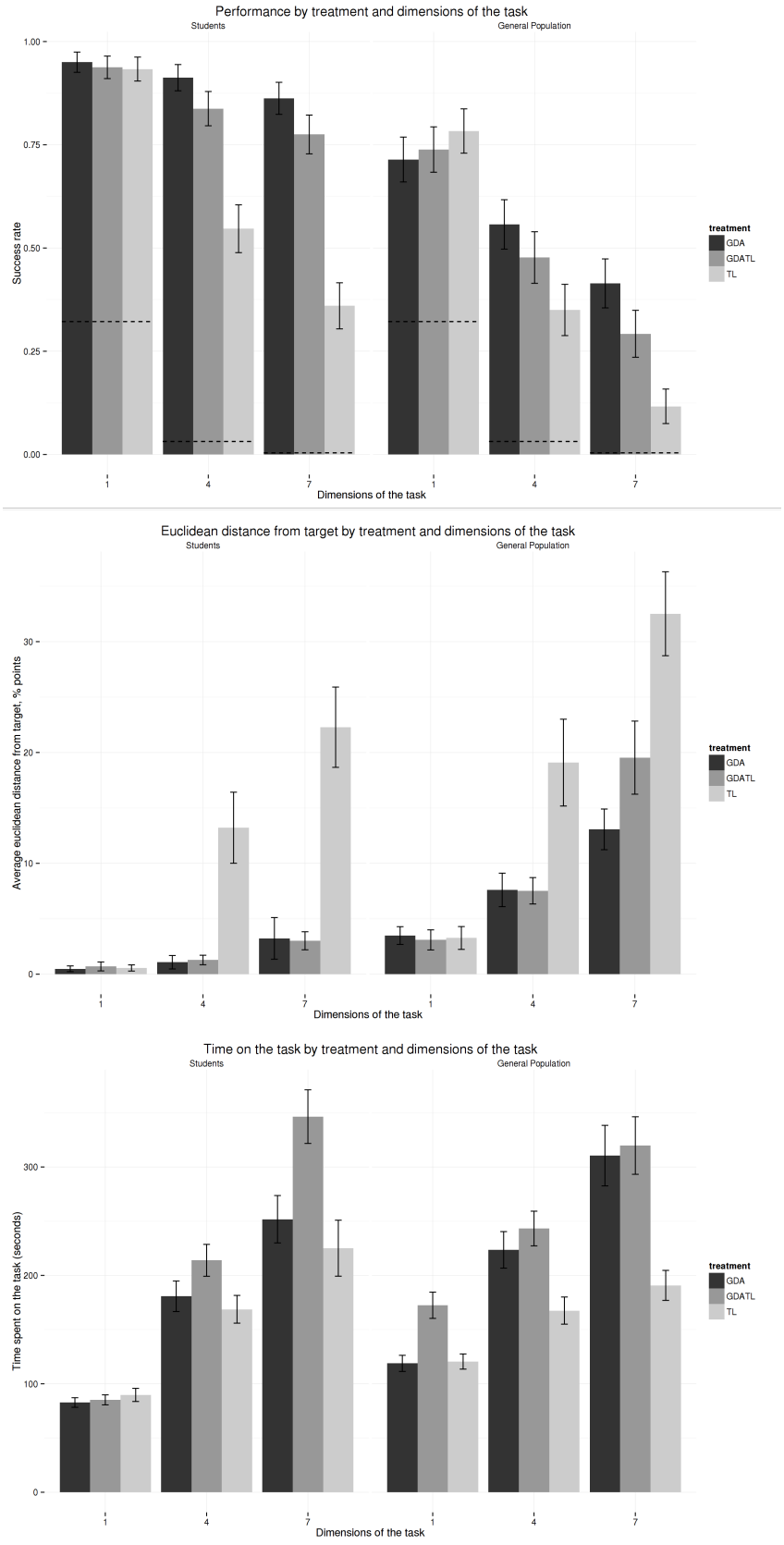


Figure 3: Performance, distance and time, by treatment, dimensions, and subject pool

treatments, as expected, since in 1D all treatments are equal. GDA leads to better performance in both 4D and 7D tasks. TL performs significantly worse than GDA for both subpopulations. GDATL, surprisingly, performs worse than GDA, even if this difference is not significant according to a Mann-Whitney test.

Success rate is a rather poor indicator of subject's behavior, since it summarises all of the subject choices in a binary variable. We get a clearer picture analyzing the distance from the target. Figure 3, middle, reports the euclidean distance from the target for each level of complexity. Results confirm by and large the previous analysis. GDATL and GDA are not significantly different. TL performs significantly worse than GDA and GDATL in both 4D and 7D tasks and over both samples. TL performs worse by a factor of roughly 6 for students, but just of 1.5 to 2.5 for the general population.

TL leads to worse performances and to less balanced diets, but it results in faster decisions, as already highlighted by the existing literature. The results in terms of time spent on each screen are summarised by Figure 3, bottom. The amount of time spent on each task is large, up to more than six minutes on average for GDATL, 7D, and increasing in task complexity. Its magnitude is comparable across the two different samples, even though students are significantly faster. TL leads to significantly faster decisions than GDATL for students, and than both GDATL and GDA for the general population.

This analysis is confirmed by a fixed effects panel regression (Table 6) that can control for unobservable heterogeneity across subjects, and includes socio-demographic characteristics and other variables from the questionnaire.[7]

Table 6 shows the results of a fixed effect logit estimation for the performance and of a linear regression model for distance.

The success rate regression shows that the effect of TL is negative for both students and the general population, though with a lower coefficient for the latter. Increasing the complexity of the tasks leads to worse performances across the board. Time spent on the task slightly increases the chances of submitting a correct answer. Among the general population, females and older subjects tend to perform worse, and higher income is correlated, with a very low coefficient, with better performance. Both indicators of heating habits – the Body Mass Index (BMI) and the question about snacking – have a positive and significant coefficient, indicating that subject having worse eating habits and being overweight perform better than average in our tasks.

The distance linear probability model results are consistent with those of success rate. In this case, though, a positive distance is a bad result, so coefficient should be interpreted in the opposite way with respect to those of the performance regression. TL has increases the distance from target, as does increasing the task complexity. Income and being female have weakly significant effects.

[7]Due to a software failure, questionnaire data for students of the TL session were not recorded. As a result, the student regressions lack the demographic controls. This is a minor nuisance since the INP students are a rather homogeneous population with respect to age, revenue and eating habits.

	General population				Students			
	Success rate		distance		% correct		distance	
	fixed effect logit		linear prob model		fixed effect logit		linear prob. model	
Constant	0.465	(0.21)	20.70	(1.31)	4.997***	(8.48)	-2.765	(-1.04)
TL	-1.281**	(-2.79)	9.403**	(2.87)	-2.619***	(-4.26)	10.42**	(2.83)
GDATL	0.0221	(0.05)	0.0125	(0.00)	-0.742	(-1.20)	0.0660	(0.02)
4 dimensions	-1.871***	(-6.56)	9.385***	(5.73)	-2.062***	(-5.54)	4.437***	(3.86)
7 dimensions	-3.172***	(-9.35)	20.80***	(11.98)	-2.800***	(-7.43)	8.654***	(7.52)
Time	0.00179*	(2.25)	-0.0207***	(-4.19)	0.00106	(1.45)	-0.0124***	(-3.53)
Controls								
Female	-0.832*	(-1.97)	6.058*	(1.98)				
Age	-0.0669***	(-3.43)	0.163	(1.19)				
Yearedu	-0.00845	(-0.10)	-0.793	(-1.23)				
Income	0.000886***	(3.31)	-0.00513**	(-2.68)				
BMI	0.135**	(2.67)	-0.548	(-1.52)				
Foodbudget	-0.00426	(-1.33)	0.0559*	(2.47)				
Snacking	1.208*	(2.57)	-4.883	(-1.45)				
$ln(\sigma_u^2)$	-0.430	(-1.02)			0.622	(1.80)		
N	585		585		705		705	

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 6: Regression analyses of Experiment 1 - % correct and mean distance from target

Experiment 2 - Time constraint

Methods

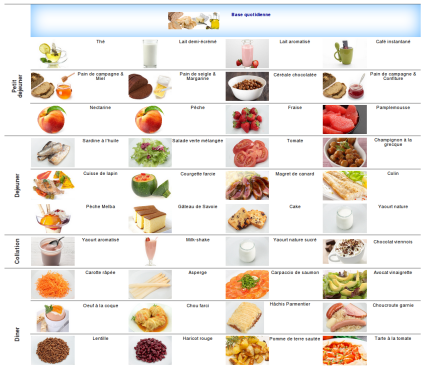
Experimental design

The aim of Experiment 2 was to gauge how time constraints affect the performance of labels. We introduced only three changes compared to Experiment 1. First, subjects were not provided with paper and pencil, and all scribbling and calculating on the side was forbidden. This constrained subjects to only use their working memory to perform computations. Second, subjects were given just 2 minutes to complete each task. This introduced stress, and the necessity to switch to fast heuristics rather than slow and time-consuming computations. Third, we introduced controls that were not present in Experiment 1 about preference, nutrition beliefs, and computing and cognitive skills. These changes allowed us to move in the direction of higher external validity – food shopping decisions are taken in the matter of seconds and do not usually involve large calculations – and to test the effect of cognitive limits on the efficiency of labels.

Procedures and materials

For the main diet-building task, we used exactly the same data and screens used in Experiment 1. We added one screen for two *control* tasks (one *preference* task and one *health* task) and two for the *mathematical* tasks.

Experiment 2 started with the two control tasks. Both tasks used image-only screens; that is a screen with the picture of the 4 food items for each of the components and no nutritional information (see Figure 4, top). In the *preference* task, subjects were asked to compose one daily diet based on their personal taste. In the following *health* task, subjects were asked with the exact same screen to compose a daily diet irrespective of their personal taste an that had to be 'as healthy as possible'. At the end of Experiment 2, we controlled for mathematical skills using two number-only screens, with no food name, no picture, but just numbers. These screens, introduced to the subjects as 'calculation exercises', where created using the same data and procedures of our standard diet screens (see Figure 4, bottom). In the 1D mathematical task subjects faced a screen displaying 12 rows of 4 numbers. Their task was to choose one number per row with the aim of keeping the sum of all the chosen numbers between 90 and 110. In the 4D math task, subjects faced for each cell of the table 4 numbers, labeled from A to D. They had to choose in order to make the sum of all numbers A lay between 90 and 110, and the sum of all numbers B, C, and D individually lower than 100. Finally, in between these controlling task, subjects went through 9 standard task screens, based on the same principles as in Experiment 1



(a) Image-Only

25			
12	5	6	0
19	29	6	20
3	3	3	2
1	3	1	3
9	8	7	4
7	9	4	10
5	6	13	11
1	2	10	1
4	15	24	4
8	10	7	7
13	5	5	6

(b) Math task, 1-dimensional

A : 25 B : 3 C : 31 D : 25			
A : 12 B : 9 C : 1 D : 3	A : 0 B : 0 C : 0 D : 0	A : 6 B : 23 C : 8 D : 4	A : 5 B : 46 C : 10 D : 4
A : 29 B : 1 C : 56 D : 15	A : 20 B : 27 C : 1 D : 23	A : 6 B : 13 C : 2 D : 6	A : 19 B : 38 C : 1 D : 23
A : 2 B : 10 C : 0 D : 0	A : 3 B : 15 C : 0 D : 0	A : 3 B : 15 C : 0 D : 0	A : 3 B : 14 C : 0 D : 0
A : 3 B : 1 C : 4 D : 17	A : 1 B : 3 C : 0 D : 0	A : 1 B : 0 C : 0 D : 0	A : 3 B : 0 C : 4 D : 6
A : 9 B : 0 C : 14 D : 4	A : 8 B : 0 C : 11 D : 3	A : 4 B : 1 C : 11 D : 21	A : 7 B : 0 C : 2 D : 6

(c) Math task, 4-dimensional

Figure 4: The mathematical control task screens

Participants

A total of 174 participants took part to Experiment 2, over 14 experimental sessions. Participants were recruited from the general population using the same newspaper ads and from the existing database of potential participants as in Experiment 1. We made sure that no subject could take part in both Experiments 1 and 2, and that no INP student participated.

The experimental sessions were run in the GAEL laboratory. The experimental software and instructions introduced only slight modifications to the ones of Experiment 1, and are available upon request.

Incentives

The *preference* task and the *health* task were not incentivized. As for the main and mathematical task, incentives were the same as in Experiment 1. The only difference was that, since the number of paid screens was reduced from 15 to 11 (9 for the main and 2 for the mathematical task), we increased the payment for each successful screen to 2.5 euro. This was done also to account for the higher opportunity cost of time of the general population with respect to the earlier student sample, and to compensate the average subject for the fact that having a reduced allotted time and no pencil and paper could result in lower overall gains. With this new incentive levels, theoretical gains could vary from 10 to 37.5 euro, for an experiment that lasted on average 1 hour and fifteen minutes. As in Experiment 1, participants were given no feedback about the success or failure of their tasks. The screens followed one another, and participants knew their performance and total earnings only at the end of the experiment.

Measures

We collected the same measures of Experiment 1. On top of those, we also recorded the success rate, distance and time spent in the added mathematical and image-only tasks.

Results

Control tasks are isomorphic with the main tasks and, hence, are directly comparable. For each task, we are able to compute the average level of each nutrient and express it in percentage of the recommended daily values. We can hence assess the distance from the recommended values when (i) subjects chose according to their personal taste (in the *preference* task), (ii) subjects chose according to their health beliefs (in the *health* task), and (iii) subjects were assisted with food labeling schemes (in the main task). Results are displayed in Figure 5, in which the shaded rectangles indicate the recommended zones.

Results show that when choosing based on their preferences (darker bars), subjects composed menus with a correct amount of calories on average, but with excessive amounts of salt (+7% with respect to the recommended daily amount), sugar (+13%) and enormous amounts of fat (+37%). The diets had enough vitamins and calcium, but lacked in fiber (-6% with respect to recommended daily amount). Simply asking the subjects to choose the healthiest possible menu (grey bars) succeeds in correcting most of the nutritional problems of the preferred diets, but only to a point. Saturated fats stayed at a level 18% higher than recommended. By and large, then, subjects could identify the healthier options in the given screen.

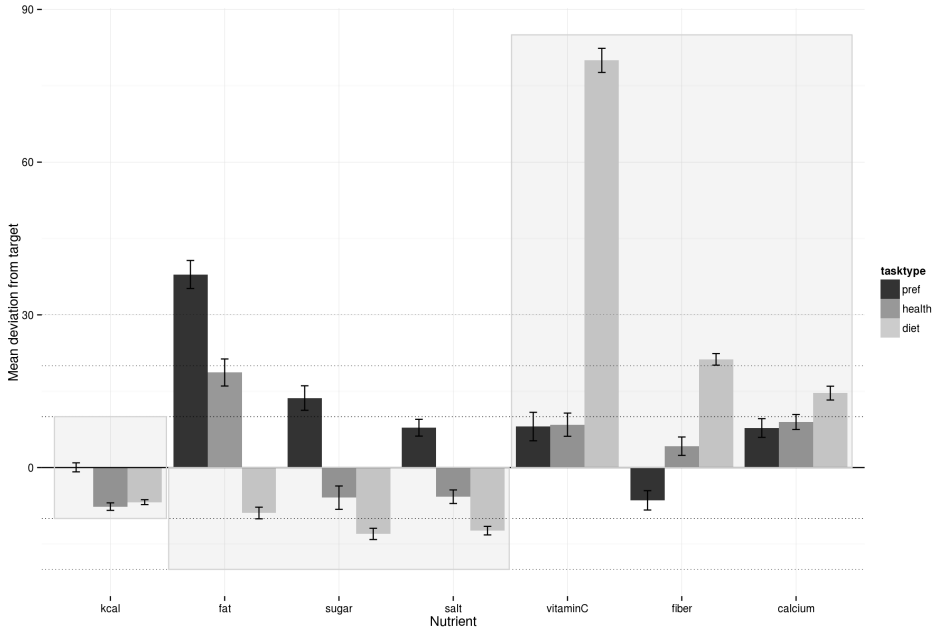


Figure 5: Mean distance from target by nutrient and task type

The light grey bars show the average over all 7D tasks, over all labels, and are meant to show the average effect on the healthiness of diets of adding *any* nutritional label, irrespective of label type. Giving the subjects nutritional information and incentivising them resulted in daily diets that were in line with the recommendations over all nutrients, including fat. Most notably, vitamin levels shot out, and the same, to a lower extent, happened to fiber. Information about vitamins might be less known than fat, sugar, and salt. Moreover, giving incentives completely crowded out preferences, and resulted in subjects choosing on average a uniformly healthy diet. Labels hence worked, over and above what was obtained by simply asking subjects to choose healthier products.

In the following we will replicate the same analysis carried out for Experiment 1. Table 7 reports the averages over treatments and dimensions of our measures: success rate, distance from target and time. The formal statistical tests over all pairwise treatment combinations for all variables are reported in Appendix AppendixB, Table B.13.

Figure 6, top, reports the average success rate by dimension and treatment. Reducing the time allotted to subjects decreases correct answers dramatically with respect to Experiment 1, and it also changes the results in terms of the best performing labels. The point estimates of the combined GDATL label are the best in both 4- and 7-dimensional tasks, even if this difference is not significant across the board, whereas its results were indistinguishable from GDA in Experiment 1. GDA and color-only TL show no difference in terms of performance, whereas TL performed much worse than GDA in Experiment 1.

Similar results are obtained in terms of distance (Figure 6, middle). The combined GDATL label performs best in both 4D and 7D tasks. In 4D tasks GDATL is the best, while in 7D it is not statistically different from TL but still significantly better than GDA.

Finally, the time spent on each screen by treatment and complexity of the task is detailed in Figure 6, bottom. The differences across treatments are much smaller than in Experiment 1, because the time limit

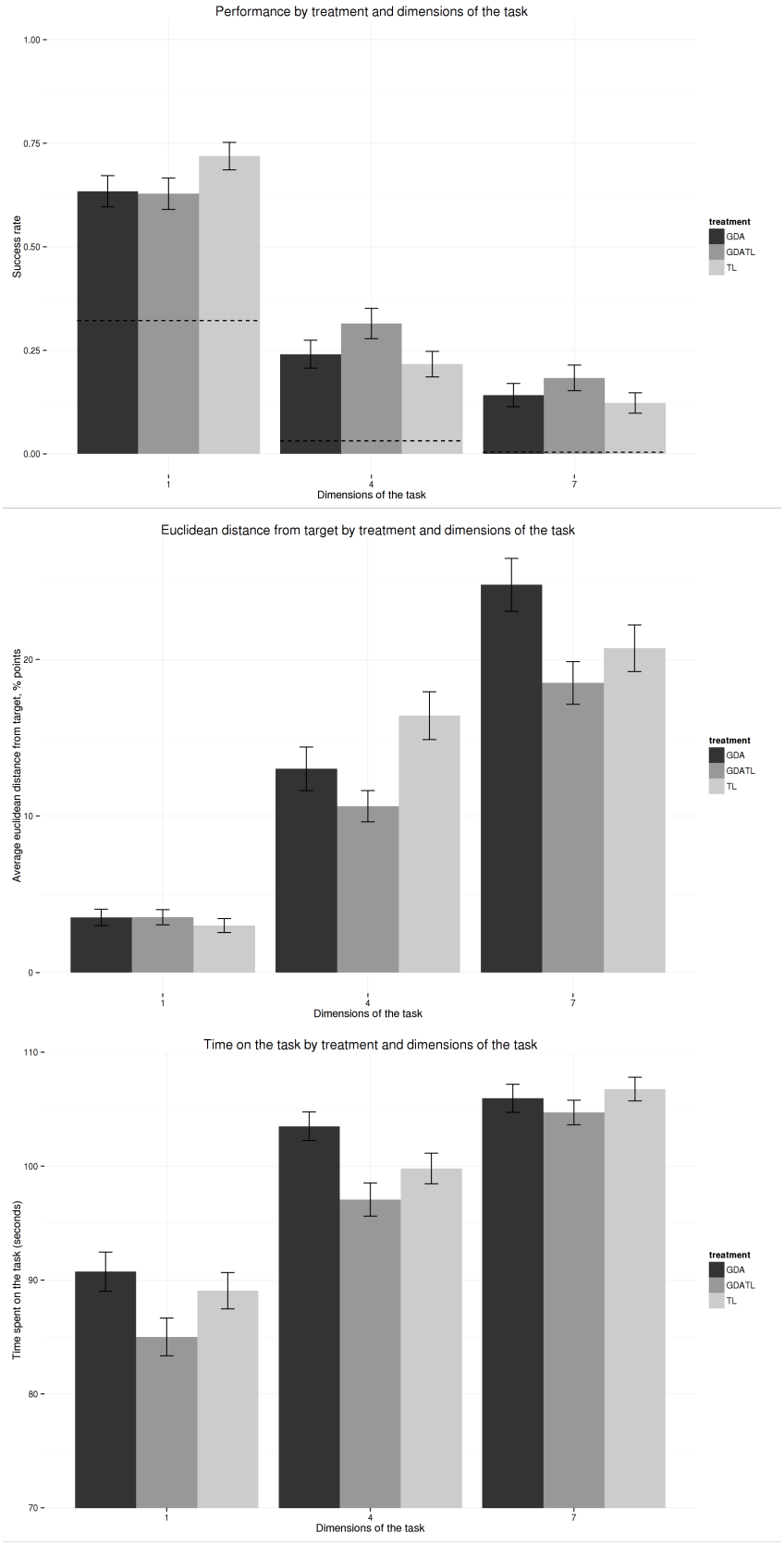


Figure 6: Average % of correct tasks, by treatment and dimensions

Dimensions	Treatment	Success rate	Distance	Time
	GDA	0.63	3.52	90.74
1	GDATL	0.63	3.54	85.01
	TL	0.72	3.01	89.06
	GDA	0.24	13.01	103.50
4	GDATL	0.31	10.63	97.06
	TL	0.22	16.41	99.79
	GDA	0.14	24.77	105.95
7	GDATL	0.18	18.50	104.70
	TL	0.12	20.71	106.76

Table 7: Experiment 2 main variables - treatment averages

compressed the right tail of the distribution. Nonetheless, GDATL leads to significantly faster decisions in 4D tasks.

Table 8 reports the result of the regression analysis. We replicate the same specifications and models of Experiment 1, running a fixed effect logit for performance and linear probability model for distance. In these regressions we include also the performance in the controls, and its interactions with the label treatments. The controls are introduced via the variables `maths`, `maths:time`, `preferences` and `health`.

The variable `maths` is computed by summing the distance from target of both 1D and 4D mathematical tasks. The distribution of this variable is rather skewed, with a large number of zeroes since many subjects did both tasks correctly, and a large variance. `maths:time` simply records the sum of the time spent on both mathematical tasks. `preferences` measures the euclidean distance from the seven targets of subjects choices in the image-only *preferences* task, while `health` is the exact same measure for the image-only *health* task. Having low scores in `preferences` means that the subject's preferences describe a healthy diet; low scores in `health` means that subjects have correct nutritional beliefs about the food items, irrespective of their preferences. The two variables have a very low correlation (0.05, not significantly different from zero), meaning that subjects' preferences are not driven by health concerns, and that subjects usually change their choices when asked to chose *healthy* food.

The results of the fixed effect logit on success rate confirm results from point estimates and statistical testing: scaling up dimensions decreases performance, but there is no statistical difference between labels. The strong result of Experiment 1 does not exist in Experiment 2. The math control shows the expected sign – being worse in the mathematical tasks reduces the likelihood of submitting correct answer, but, interestingly, not so for TL or GDATL tasks. Individual preferences have no significant impact on success rate or distance – a result that confirms that our incentives might have crowded out preferences completely. As in Experiment 1, females tend to submit worse diets and age and education show similar patterns.

The results of the distance linear probability model show a very similar pattern. When taking into account distance, though, GDATL results in significantly reduced distance with respect to the base GDA case. `math` is significant, but only for GDA and not for TL and GDATL treatments, indicating that

	Success rate		distance	
	fixed effect logit		linear prob. model	
Constant	0.309	(0.22)	-3.549	(-0.46)
TL	0.536	(0.98)	-5.643	(-1.85)
GDATL	0.911	(1.57)	-6.409*	(-1.97)
4 dimensions	-1.990***	(-12.31)	7.536***	(5.98)
7 dimensions	-3.757***	(-17.56)	29.74***	(22.93)
Time	-0.00132	(-0.43)	0.0841***	(4.70)
Performance in control tasks				
maths	-0.0458***	(-3.70)	0.144*	(2.47)
maths×TL	0.0109	(0.61)	-0.0214	(-0.29)
maths×GDATL	0.00604	(0.36)	0.0660	(0.80)
maths:time	0.00127	(0.39)	-0.0228	(-1.27)
preferences	0.00665	(1.21)	0.0396	(1.30)
preferences×TL	-0.00738	(-0.96)	0.0669	(1.55)
preferences×GDATL	-0.00591	(-0.76)	0.0405	(0.93)
health	-0.000796	(-0.25)	0.0962***	(5.30)
Controls				
Female	-0.623**	(-3.11)	0.819	(0.73)
Age	-0.0177*	(-2.00)	0.0464	(0.94)
Yearedu	0.108*	(2.25)	-0.418	(-1.57)
Income	0.00000592	(0.06)	-0.000447	(-0.76)
BMI	0.0149	(0.78)	0.0270	(0.26)
Foodbudget	-0.00144	(-0.84)	-0.00142	(-0.15)
Snacking	-0.0417	(-0.18)	1.711	(1.31)
$\ln(\sigma_u^2)$	-0.783*	(-2.50)		
N	1846		1802	

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 8: Regression analysis of Experiment 2 – Success rate and distance form target

subjects relied on other decision heuristics than outright computation when faced with color codes. According to intuition, worse performance in the health task is correlated to worse distance in the main task. No demographic variable has a significant impact on distance in Experiment 2.

Discussion and conclusion

The research presented in this paper stems from a basic question: which labelling scheme between Guideline Daily Amounts (GDA) and Traffic Lights (TL) is the better tool to assist consumers willing to consume healthy diets? By providing all the necessary information to compute the recommended daily amount, GDA seems better equipped than the coarser TL. Nevertheless, such computations require effort and skill, in particular when the nutritional goals are multiple. On the other hand, TL are salient and intuitive, allowing the use of simpler heuristics.

In order to address this research question, we built an artificial experimental setting that abstracts away from exposure, preference and actual choice (*i.e.* purchase). With aligned incentives, and full focus, subjects are asked to build daily diets that satisfy pre-determined nutritional goals. We provide an objective and controlled benchmark to test the efficacy of GDA, TL and the combined GDATL with respect to their stated goal of helping consumers build a healthy diet.

With no time constraint, Experiment 1 shows that GDA and GDATL outperform TL, irrespective of the number of nutritional goals that are simultaneously set to the subjects. Nonetheless, this result is fragile. Experiment 2 shows that setting a reasonable time constraint is enough to invalidate the previous result: TL and GDA have identical performance with 4 nutritional goals and TL even outperforms GDA with 7 nutritional goals.

Computations are key to understand these results. GDA is the right tool for decision makers that decide to compute their way out of the given task. With unlimited time and computational tools (*i.e.* pen and paper), the setting of Experiment 1 was ideal for GDA. The most widespread strategy in Experiment 1 was indeed to take as much time as possible, and run through all the computations. Engineering students, highly skilled in mathematics, performed about twice as good as the general population. In Experiment 2, the mathematical skills as recorded in the additional control tasks were correlated with higher success rates in GDA screens. The absence of correlation with TL and GDATL screens shows that subjects switched to heuristics other than plain computation in the presence of colour-codes in Experiment 2.

This paper contributes to the GDA and TL debate from a different perspective than the one usually adopted in the existing literature. Most studies assessing the efficacy of GDA and TL rely on choice experiments based on a very limited set of products (2 or 3 items). We instead focus on the ability of consumers to build healthy diet irrespective of their personal taste in the cleanest possible environment.

This rather artificial setting very possibly limits the direct applicability of our results to the real world. First, our environment, unlike the real world, is highly favourable to computations and, therefore, arguably best suited for GDA. Nonetheless, this choice has the advantage of enabling us to identify the upper limit of efficacy for GDA and at the same time observe how the cognitive and skill limitations of consumers might affect its performance. If GDA fails in our biased environment, with incentives and focalised attention, it cannot excel in the noisier, fuzzier, complicated real world. Second, we do not

address the question of the labels' impact on consumers' preferences. In our experiments there is no actual food choice, since consumers do not buy nor will eat any of the products. One may speculate, for instance, that in real food purchases, more salient labels like TL would be better equipped to change consumers' behaviours. But such assertions will remain intuitions as far as they are not properly tested. By providing strong evidence on the functioning of the labels *in vitro*, our paper sets the ground for further research on the integration of label information in actual consumer decision processes.

References

Borgmeier, I. and Westenhoefer, J. (2009). Impact of different food label formats on healthiness evaluation and food choice of consumers: a randomized-controlled study. *BMC Public Health*, 9.

Darmon, N., Vieux, F., Maillot, M., Volatier, J., and Martin, A. (2009). Nutrient profiles discriminate foods according to their contribution to nutritionally adequate diets: a validation study using linear programming and the sain, lim system. *American Journal of Clinical Nutrition*, 89(4):1227–1236.

Drichoutis, A. C., Nayga, R. M., and Lazaridis, P. (2011). *Nutritional Labeling*, chapter 20, pages 520–545. Oxford University Press, Oxford.

FSA (2013). Guide to creating a front of pack (fop) nutrition label for pre-packed products sold through retail outlets. Guidance of the Department of Health of the United Kingdom of Great Britain and Northern Ireland.

Greifeneder, R., Scheibehenne, B., and Kleber, N. (2010). Less may be more when choosing is difficult: Choice complexity and too much choice. *Acta Psychologica*, 133(1):45 – 50.

Grunert, K. G. and Wills, J. M. (2007). A review of european research on consumer response to nutrition information on food labels. *Journal of Public Health*, 15(5):385–399.

Grunert, K. G., Wills, J. M., and L., F.-C. (2010). Nutrition knowledge, and use and understanding of nutrition information on food labels among consumers in the uk. *Appetite*, 55(2):177–189.

Hawley, K. L., Roberto, C. A., Bragg, M. A., Liu, P. J., Schwartz, M. B., and Brownell, K. D. (2013). The science on front-of-package food labels. *Public Healthealth Nutrition*, 16:430–9.

INSERM (2006). *Su-Vi-Max: Table de composition des aliments*. Paris: Economica.

Kelly, B., Hughes, C., Chapman, K., Louie, J. C.-Y., Dixon, H., Crawford, J., King, L., Daube, M., and Slevin, T. (2009). Consumer testing of the acceptability and effectiveness of front-of-pack food labelling systems for the australian grocery market. *Health Promotion International*, 24(2):120–129.

Malhotra, N. K. (1982). Information load and consumer decision making. *Journal of consumer research*, pages 419–430.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological review*, 63(2):81.

Moeser, A., Hoefkens, C., Camp, J., and Verbeke, W. (2010). Simplified nutrient labelling: consumers' perceptions in germany and belgium. *Journal fuer Verbraucherschutz und Lebensmittelsicherheit*, 5(2):169–180.

Synovate, on behalf of the Central Office of Information, F. S. A. (2005). Quantitative evaluation of alternative food signposting concepts: Report of findings. Technical report, West Malling, UK: Synovate.

Union, E. (2011). Regulation no 1169/2011. EU Official Bulletin 22/11/2011 L304/61 Annex XIII.

Vyth, E. L., Steenhuis, I. H. M., Brandt, H. E., Roodenburg, A. J. C., Brug, J., and Seidell, J. C. (2012). Methodological quality of front-of-pack labeling studies: a review plus identification of research challenges. *Nutrition reviews*, 70:709–20.

Nutrient	GDA	Unit
Energy	2000	kcal
Fat	70	gram
Saturates	20	gram
Sugars	90	gram
Sodium	2.4	gram
Fiber	25	gram
Vitamin C	80	milligram
Calcium	800	milligram

Table A.9: Reference table for conversion of nutritional information into GDA

AppendixA. Nutritional details of GDA and TL

We built the GDA and TL data using official sources. GDA was computed from nutritional data applying the official regulations of the European Union (2011) Official Bulletin detailed in Table A.9

TL thresholds were computed applying a simplified version of the official TL guide of the UK FSA (2013). The actual thresholds in terms of GDA are reported in Table A.10. Note that the actual threshold do not differ substantially from the 5-25 simplified threshold employed, and that no thresholds are provided for the *good* nutrients; in this case we implemented our simple 5-25 rule, inverted.

	Sugar	Salt	Fat	Vitamin C	Fiber	Calcium
Green	$< 5.55\%$	$< 5\%$	$< 4.28\%$	-	-	-
Red	$> 16.6\%$	$> 25\%$	$> 28.57\%$	-	-	-

Table A.10: Reference thresholds to convert GDA into TL

AppendixB. Complete statistical tests

Variable	Dimensions	Test	Mean GDA	Mean GDATL	Mean TL	MW p-value
Success rate	1	GDATL v GDA	0.95	0.94		0.736
		TL v GDA	0.95		0.93	0.662
		TL v GDATL		0.94	0.93	0.92
	4	GDATL v GDA	0.91	0.84		0.154
		TL v GDA	0.91		0.55	**0**
		TL v GDATL		0.84	0.55	**0**
	7	GDATL v GDA	0.86	0.78		0.153
		TL v GDA	0.86		0.36	**0**
		TL v GDATL		0.78	0.36	**0**
Distance	1	GDATL v GDA	0.49	0.70		0.745
		TL v GDA	0.49		0.56	0.668
		TL v GDATL		0.70	0.56	0.923
	4	GDATL v GDA	1.08	1.28		0.159
		TL v GDA	1.08		13.21	**0**
		TL v GDATL		1.28	13.21	**0**
	7	GDATL v GDA	3.23	3.01		0.137
		TL v GDA	3.23		22.28	**0**
		TL v GDATL		3.01	22.28	**0**
Time	1	GDATL v GDA	82.77	85.19		0.591
		TL v GDA	82.77		89.77	0.759
		TL v GDATL		85.19	89.77	0.984
	4	GDATL v GDA	180.76	213.91		**0.03**
		TL v GDA	180.76		168.74	0.944
		TL v GDATL		213.91	168.74	**0.01**
	7	GDATL v GDA	251.81	346.44		**0.002**
		TL v GDA	251.81		225.10	0.162
		TL v GDATL		346.44	225.10	**0**

Table B.11: Mann-Whitney Rank-Sum tests for Experiment 1 - INP Students

Variable	Dimensions	Test	Mean GDA	Mean GDATL	Mean TL	MW p-value
Success rate	1	GDATL v GDA	0.71	0.74		0.756
		GDATL v TL		0.74	0.78	0.561
		GDA v TL	0.71		0.78	0.371
	4	GDATL v GDA	0.56	0.48		0.354
		GDATL v TL		0.48	0.35	0.153
		GDA v TL	0.56		0.35	**0.019**
	7	GDATL v GDA	0.41	0.29		0.141
		GDATL v TL		0.29	0.12	**0.016**
		GDA v TL	0.41		0.12	**0**
Distance	1	GDATL v GDA	3.49	3.09		0.682
		GDATL v TL		3.09	3.27	0.658
		GDA v TL	3.49		3.27	0.42
	4	GDATL v GDA	7.60	7.52		0.523
		GDATL v TL		7.52	19.09	0.078
		GDA v TL	7.60		19.09	**0.013**
	7	GDATL v GDA	13.06	19.54		0.232
		GDATL v TL		19.54	32.52	**0.002**
		GDA v TL	13.06		32.52	**0**
Time	1	GDATL v GDA	118.85	172.46		**0**
		GDATL v TL		172.46	120.55	**0.001**
		GDA v TL	118.85		120.55	0.714
	4	GDATL v GDA	223.54	243.22		0.184
		GDATL v TL		243.22	167.55	**0**
		GDA v TL	223.54		167.55	**0.025**
	7	GDATL v GDA	310.47	319.77		0.493
		GDATL v TL		319.77	190.78	**0**
		GDA v TL	310.47		190.78	**0.003**

Table B.12: Mann-Whitney Rank-Sum tests for Experiment 1 - General Population

Variable	Dimensions	Test	Mean GDA	Mean GDATL	Mean TL	MW p-value
Success rate	1	GDA v GDATL	0.63	0.63		0.91
		TL v GDA	0.63		0.72	0.091
		TL v GDATL		0.63	0.72	0.071
	4	GDA v GDATL	0.24	0.31		0.137
		TL v GDA	0.24		0.22	0.597
		TL v GDATL		0.31	0.22	**0.04**
	7	GDA v GDATL	0.14	0.18		0.32
		TL v GDA	0.14		0.12	0.609
		TL v GDATL		0.18	0.12	0.122
Distance	1	GDA v GDATL	3.52	3.54		0.845
		TL v GDA	3.52		3.01	0.174
		TL v GDATL		3.54	3.01	0.12
	4	GDA v GDATL	13.01	10.63		0.412
		TL v GDA	13.01		16.41	0.145
		TL v GDATL		10.63	16.41	**0.021**
	7	GDA v GDATL	24.77	18.50		**0.008**
		TL v GDA	24.77		20.71	**0.042**
		TL v GDATL		18.50	20.71	0.365
Time	1	GDA v GDATL	90.74	85.01		**0.015**
		TL v GDA	90.74		89.06	0.416
		TL v GDATL		85.01	89.06	0.085
	4	GDA v GDATL	103.50	97.06		**0.001**
		TL v GDA	103.50		99.79	0.054
		TL v GDATL		97.06	99.79	0.141
	7	GDA v GDATL	105.95	104.70		0.058
		TL v GDA	105.95		106.76	0.941
		TL v GDATL		104.70	106.76	**0.035**

Table B.13: Mann-Whitney Rank-Sum tests for Experiment 2