Journal of Economic Psychology 55 (2016) 30–50

Contents lists available at ScienceDirect

Journal of Economic Psychology

journal homepage: www.elsevier.com/locate/joep

Helping consumers with a front-of-pack label: Numbers or colors? ☆
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

ARTICLE INFO

Article history:
Received 30 June 2015
Received in revised form 19 February 2016
Accepted 22 March 2016
Available online 24 March 2016

JEL classification:
D12
D18
C91
C93

PsycINFO classification:
3920

Keywords:
Nutritional labels
Food choice
Experimental economics
Guideline Daily Amount
Reference intake
Traffic Lights

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 labeling schemes in assisting consumers to build a healthy daily menu. Participants must compose a daily menu, choosing from a finite set of products, and are paid a fixed cash amount only if the menu satisfies pre-determined nutritional goals. Goals correspond to achieving the Guideline Daily Amount values of 1 (kcal), 4 (kcal, fat, sugar, salt) or 7 (kcal, fat, sugar, salt, fiber, vitamin C and calcium) different nutritional attributes. 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.
© 2016 Elsevier B.V. All rights reserved.

☆ 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, Paris and Rennes. All remaining errors are ours.
* Corresponding author.
E-mail addresses: paolo.crosetto@grenoble.inra.fr (P. Crosetto), laurent.muller@grenoble.inra.fr (L. Muller), bernard.ruffieux@grenoble-inp.fr (B. Ruffieux).

http://dx.doi.org/10.1016/j.joep.2016.03.006
0167-4870/© 2016 Elsevier B.V. All rights reserved.

1. 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, Nayga, & Lazaridis, 2011, chap. 20), the debate over their relative efficacy mainly focuses on two widespread labeling 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, per serving or per 100 grams. While GDA express 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 have been recently replaced by *Reference Intake*[1] 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, patronizing 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 & 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 menu. 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 behavior.

Many studies have fueled the debate about what format between GDA and TL should be favored (see Kelly et al. (2009), Moeser, Hoefkens, Camp, & Verbeke (2010), Grunert et al. (2010) for experimental studies and Grunert & 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 three products according 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 products 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 labeling schemes. Questions seem to favor 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, Borgmeier & Westenhoefer, 2009, Kelly et al., 2009). On the other hand, questions seem to favor 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 maximize the amount of good nutrients, namely vitamin C, calcium and fiber.

The experiment was designed so that GDA and TL are only assessed on their efficacy to help consumers build a healthy daily menu. We abstract away from considerations 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

[1] 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.

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 favors 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, Scheibehenne, & Kleber, 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 menu 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 min 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 min 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 and subjects' (i) preferences and (ii) nutritional 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. Controlling for subject's preferences and performance in the mathematical and preference-based tasks, the effect is muted, the three labels performing equally. 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 all three labels have the same performance once mathematical abilities and food preferences are controlled for.

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 does not excel in a setting strongly biased in its favor, then it might only do worse in the case of real purchases, in which consumers are severely bounded in time, attention, focus, and budget.

2. Experiment 1 – benchmark

2.1. Methods

2.1.1. Experimental design and treatments

The aim of the first experiment was to build a controlled environment to test the relative efficacy of three different labeling 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 were asked to compose *daily menus* that satisfied a set of pre-determined nutritional criteria. We imposed a structure to the *daily menus* that our subject had to build. Daily menus were divided quite naturally (to a French/continental consumer) into four meals, that is, breakfast, lunch, afternoon snack and dinner. Each meal was composed by one or more courses. Over the four meals, a total of 11 courses were available for the daily menu composition. A twelfth course representing the food daily basis (i.e. bread, oil, butter) remained fixed and common to all participants. For the remaining 11 courses, subjects had to choose one among 4 available food products.

A treatment-specific nutritional label was provided for each product. Subjects knew from the outset that they had to repeat this task 15 times, with different food products to choose from on each repetition, and that they would be paid €1.50 for each successfully completed task (i.e., if the composed daily menu 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), on top of the caloric constraint, bad nutrients had to be kept below a pre-determined limit. In seven-dimensional tasks (7D), on top of the 4D constraints, 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.

Each task was displayed on a single screen. One screen contained a row for the daily basis of bread, oil, and butter, and, for each of the 11 remaining courses that compose a daily menu, a row with 4 alternative food products. That is, a screen was a 12-row, 4-column table, and subject had to choose one cell – one product – for each (but one) row. We used 5 different screens that included different products. Each screen was repeated three times, once with 1-dimensional information, once in 4D and once in 7D. The position of the products 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 Fig. 1. Details on the products and the procedure used to create our screens are given in the next section.

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.

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.

2.1.2. Procedures and materials

We gathered full nutritional data for 346 widely consumed general products. The nutritional data came from the Su-Vi-Max project (INSERM, 2006)[2] complemented with data from the Darmon, Vieux, Maillot, Volatier, and Martin (2009) study. We coded each product 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 product is represented by its ready-to-eat picture, centered on a neutral white background. 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.

Recommended daily amounts were generated from the raw nutritional data following European Union (2011) official tables, with reference to an average adult.[3] 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 to our participants, with no substantial differences with reality.[4] The used threshold are summarized in Table 2.

We used our product database to build possible daily menus, i.e., lists of products that make up for a full day of consumption. The daily menus were built with reference to traditional French eating habits, and were composed of 12 courses: a light continental breakfast, lunch (starter, main course, dessert), afternoon snack, and dinner (starter, main course, side and dessert). To balance the menus with commonly used products such as bread, oil or butter, we added a *daily base* composed of 120 grams of white bread, 20 grams of oil and 10 of butter. A daily menu was hence overall composed of 11 courses and a common base. Examples of which products were assigned to each meal are given in Table 3.

We randomly generated several thousand daily menus, and then checked them against nutritional criteria, with the aim of singling out both healthy and unhealthy menus. We considered a daily menu healthy if it simultaneously satisfied all 7 nutritional goals. That is, a daily menu was deemed healthy if it contained, over the 12 courses, 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 menus unhealthy if they failed all seven criteria. After checking the menus for abnormal products (i.e., products that would seem out of place for French eating habits), we selected 12 healthy and 12 unhealthy menus. Participants faced decision screens made of four daily menus, 2 healthy and 2 unhealthy. This generated screens made up of four columns (the four daily menus) and 12 rows (the courses of a daily menu).

The information content of the screens varied across treatments. In the GDA treatment, all information was displayed as percentage 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 percentage of the recommended daily amount and never as color 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 products (see Fig. 2)).

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

[3] The full conversion table is available in Appendix A.

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

Table 1
Constraints given to subjects for daily menus, by dimension.

Dimensions	Nutrient	Threshold (as % of recommended daily amount)
1D	kcal	$90 \leqslant \sum$ Kcal $\leqslant 110$
4D	Sugar	$\sum$ Sugar $\leqslant 100$
	Saturated fat	$\sum$ Fat $\leqslant 100$
	Salt	$\sum$ Salt $\leqslant 100$
7D	Fiber	$\sum$ Fiber $\geqslant 100$
	Vitamin C	$\sum$ VitaminC $\geqslant 100$
	Calcium	$\sum$ Calcium $\geqslant 100$

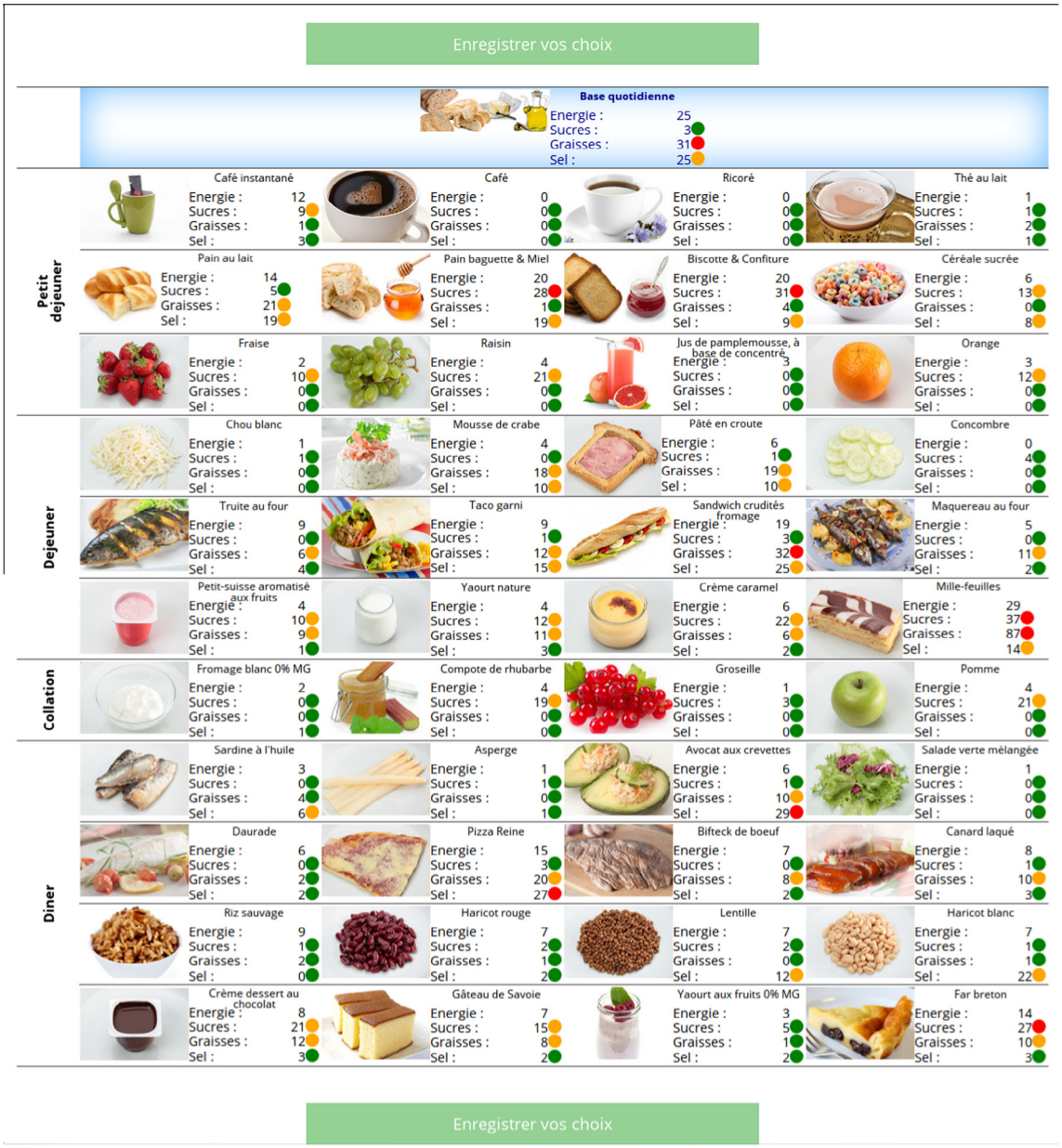


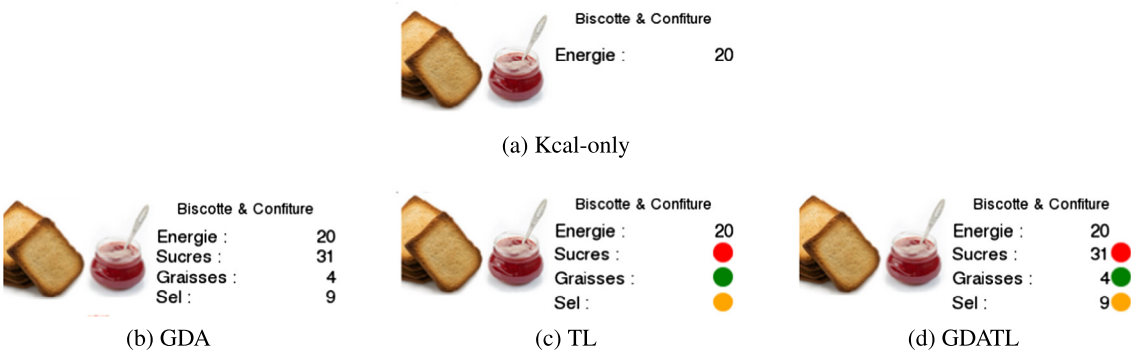
Fig. 1. Screenshot of a 4D, GDATL decision screen.

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

	% 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 3
The chosen structure of a daily menu, with example products.

Daily base	–	120 g bread, 10 g butter, 20 g 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...



(a) Kcal-only (b) GDA (c) TL (d) GDATL

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

We generated all the possible screens from our 24 starting daily menus, 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 repeated 100.000 times a random choice from each of 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 menus 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).

Table 4
Average performance of the 100.000 random players on the final screens.

	Probability (%) of satisfying the constraint when playing randomly								
	kcal	*Fat*	*Sugar*	*Salt*	*Fiber*	*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

2.1.3. 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 center 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 in Appendix C.

2.1.4. 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, in a pattern of 1D, 4D, 7D, repeated 5 times, and participants knew their total earnings only at the end of the experiment.

2.2. 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 menu 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 compute 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% to 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 measure the time spent on each task by each subject, in milliseconds.

Questionnaire. At the end of the experiment subjects filled a short questionnaire. It 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.

2.3. Results

The average results of our measures by subject pool, label and dimensions are displayed in Table 5. The formal statistical tests of hypotheses, carried out using the χ^2 test for success rate and the Mann–Whitney non-parametric test for distance and time spent on the task, are reported in detail for all pairs of treatments in Appendix B, in Table 11 for INP students and in Table 12 for the general population. Differences between students and general population are significant across the board (not reported).

Results are also reported in graphical form. The treatment means of the performance in the task is reported in Fig. 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.

Table 5
Experiment 1 main variables – treatment averages.

Dimensions	Treatment	Students			General population		
		Success rate	Distance	Time (s)	Success rate	Distance	Time (s)
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

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 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 χ^2 test.

Success rate is a rather poor indicator of subject's behavior, since it summarizes all of the subject choices in a binary variable. We get a clearer picture analyzing the distance from the target. Fig. 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 1.5–2.5 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 summarized by Fig. 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 panel regressions with standard errors clustered at the individual level. Table 6 reports marginal effects of the random effect logit estimation of success rate and coefficients of the linear regression model of distance from target. The regressions include the interactions of treatment and dimension variables as well as socio-demographic characteristics and other variables from the questionnaire.[5]

The success rate regression shows that increasing the complexity of the tasks leads to worse performances for all labels, but especially so for the general population and for TL, while GDA and GDATL have similar effects. Among students, success rates in the 1- and 4-dimensional cases are not different for GDA and GDATL, while TL has a negative impact. Among the general population, females and older subjects tend to perform worse, while higher income is correlated 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 regression model results are consistent with those of success rate. In this case, though, a positive distance is a bad result, so coefficients should be interpreted in the opposite way with respect to those of the performance regression. TL performs worse than GDA and GDATL in 7-dimensional tasks. Spending more time on the task decreases the distance from the target. Controls coefficient are consistent with the results of the success rate regressions.

3. Experiment 2 – time constraint

3.1. Methods

3.1.1. 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 min to complete each task. This introduced stress, and the necessity to switch to fast heuristics

[5] 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.

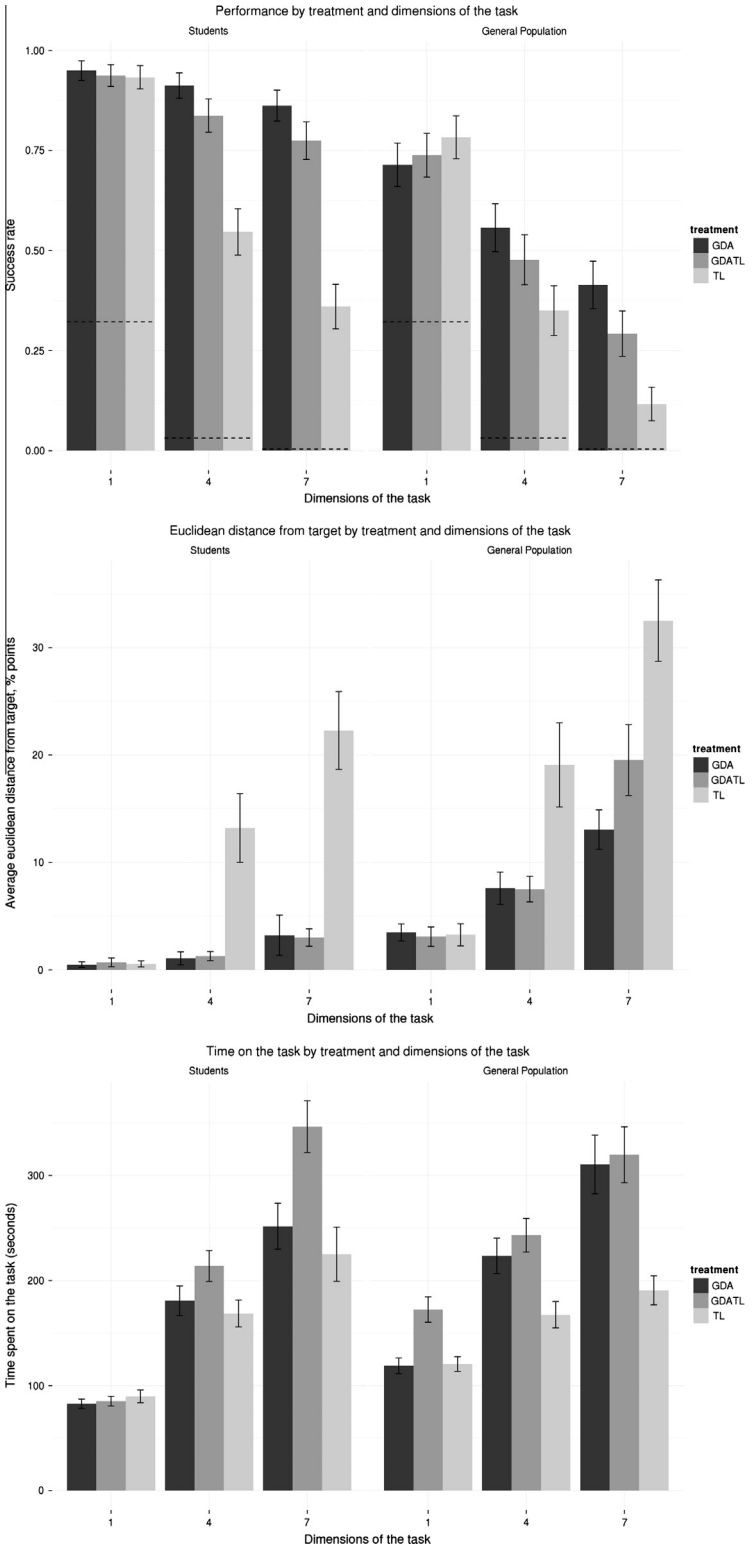


Fig. 3. Performance, distance and time, by treatment, dimensions, and subject pool.

Table 6
Regression analyses of Experiment 1 – success rate and euclidean distance from target.

	General population				Students			
	Success rate random effect logit		Distance linear regression		Success rate random effect logit		Distance linear regression	
Constant			24.17	(1.29)			1.269*	(1.97)
TL	0.0360	(0.38)	0.169	(0.06)	−0.0310	(−0.34)	0.139	(0.32)
GDATL	0.0972	(1.11)	−2.051	(−0.79)	−0.0345	(−0.38)	0.235	(0.43)
4 dimensions	−0.167*	(−2.55)	5.940***	(3.42)	−0.0789	(−1.15)	1.516	(1.37)
7 dimensions	−0.315***	(−4.62)	12.92***	(4.26)	−0.146*	(−2.13)	4.338	(1.23)
Time (minutes)	0.0129	(1.77)	−1.045*	(−2.24)	0.00470	(1.09)	−0.566	(−1.51)
Interactions								
TL×4D	−0.301**	(−2.82)	10.70	(1.78)	−0.238*	(−2.56)	11.88	(1.83)
TL×7D	−0.458***	(−3.76)	17.56**	(2.88)	−0.278**	(−3.09)	18.66*	(2.45)
GDATL×4D	−0.0981	(−1.02)	−0.276	(−0.14)	−0.0512	(−0.58)	0.279	(0.42)
GDATL×7D	−0.145	(−1.43)	6.094	(1.46)	−0.0424	(−0.50)	0.442	(0.19)
Controls								
Female	−0.150*	(−2.47)	5.939**	(2.90)				
Age	−0.0102***	(−3.92)	0.164	(1.10)				
Year Education	−0.00323	(−0.25)	−0.800	(−1.30)				
Income ('000€/month)	0.139***	(3.81)	−5.176***	(−3.58)				
BMI	0.0237**	(3.23)	−0.549*	(−2.11)				
Foodbudget (€)	−0.000751	(−1.62)	0.0551***	(3.36)				
Snacking	0.205**	(3.05)	−4.973**	(−2.85)				
n subjects	39		39		47		47	
N	585		585		705		705	

t statistics in parentheses.
For logit estimations marginal effects at the mean are reported.
*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

rather than slow and time-consuming computations. Third, we introduced controls that were not present in Experiment 1 about preferences, 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.

3.1.2. Procedures and materials

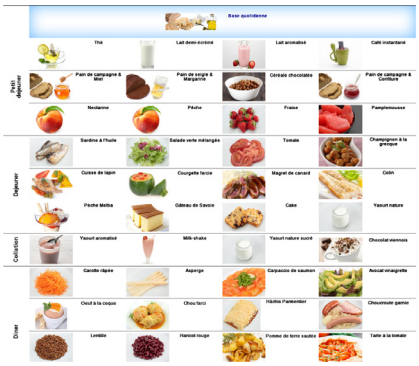
For the main diet-building task, we used exactly the same data and screens used in Experiment 1. We added one screen for two *control* tasks (one *preference* task and one *health* task) and two for the *mathematical* tasks.

Experiment 2 started with the two control tasks. Both tasks used image-only screens; that is a screen with the picture of the 4 products for each of the courses and no nutritional information (see Fig. 4, top). In the *preference* task, subjects were asked to compose one daily menu based on their personal taste. In the following *health* task, subjects were asked with the exact same screen to compose a daily menu 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 menu-building screens (see Fig. 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.

3.1.3. 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. The instructions are reported in Appendix C.



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

(c) Math task, 4-dimensional

Fig. 4. The mathematical control task screens.

3.1.4. Incentives

The *preference* task and the *health* task were not incentivized.[6]

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 h 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, in three sequences of one 1D, one 4D, and one 7D screen, and participants learned their performance and total earnings only at the end of the experiment.

3.2. 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 (preference and health) tasks.

3.3. 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 Fig. 5, in which the shaded rectangles indicate the recommended zones.

[6] The *preference* task could have been incentivized by means of selling some of the products, but this would have been unpractical, and required training and instructions; for the *health* task it would have been impossible to provide transparent incentives, since the objective was vaguely defined. Besides, providing incentives and instructions for these tasks would have diluted the incentive of the main task. Since we used them only as controls, we deemed better to have no incentives.

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 (gray 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. The light gray 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 menus 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 (χ^2 and Mann–Whitney rank-sum) over all pairwise treatment combinations for all variables are reported in Appendix B, Table 13.

Fig. 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 (Fig. 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 Fig. 6, bottom. The differences across treatments are much smaller than in Experiment 1, because the time limit compressed the right tail of the distribution. Nonetheless, GDATL leads to significantly faster decisions in 4D tasks.

Table 8 reports the results of the regression analysis. We replicate the same specifications and models of Experiment 1, running a random effect logit for performance and linear regression model for distance, clustering standard errors at the individual level. We report directly marginal effects at the mean for the logit regression. 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 products, 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 choose *healthy* food.

The results of the random 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 (no significant interaction coefficient). 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, and this result is not different across treatments. 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. Income is no more significant, and education is only marginally significant, possibly indicating that the *math* is a mediator for the effects.

The results of the distance linear regression model show a very similar pattern: the complexity of the task matters, but labels are not different among each other at each complexity level. As in Experiment 1, time spent on the task has a significant impact.*math* is significant and has the expected sign. 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.

4. Discussion and conclusion

The research presented in this paper stems from a basic question: which labeling 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.

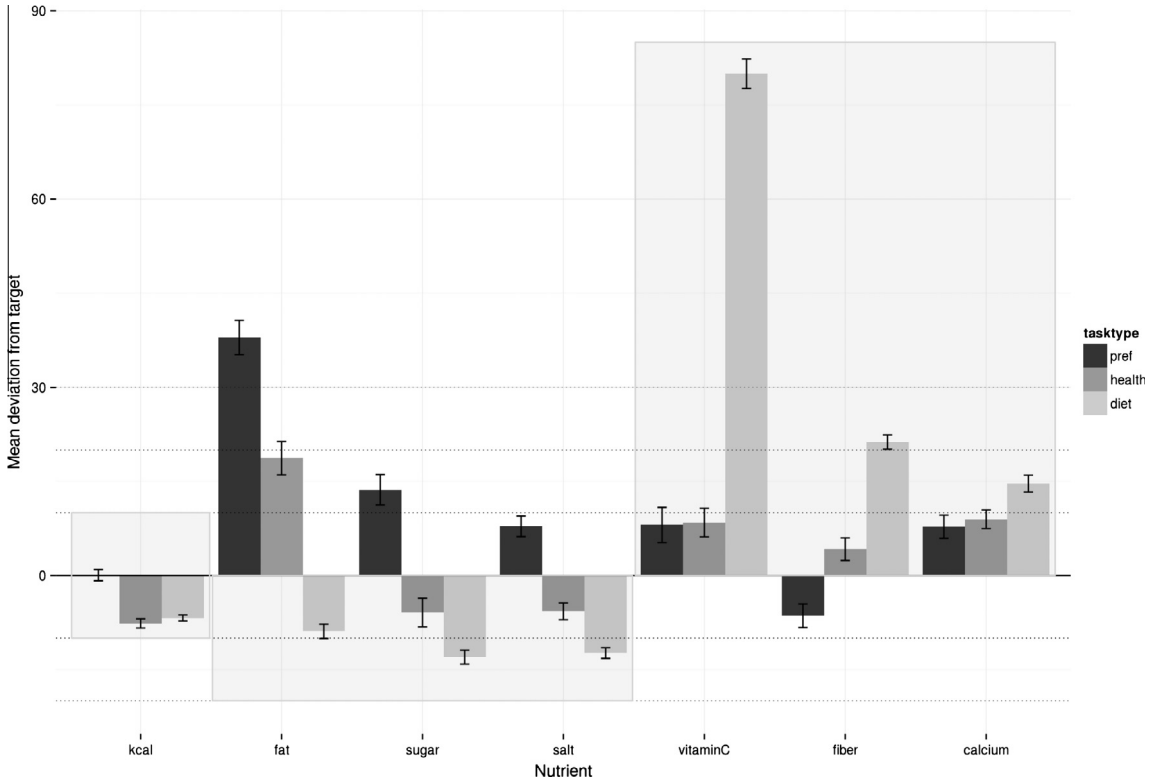


Fig. 5. Mean distance from target by nutrient and task type.

Table 7
Experiment 2 main variables – treatment averages.

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

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 menus 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 and 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 across all treatments.

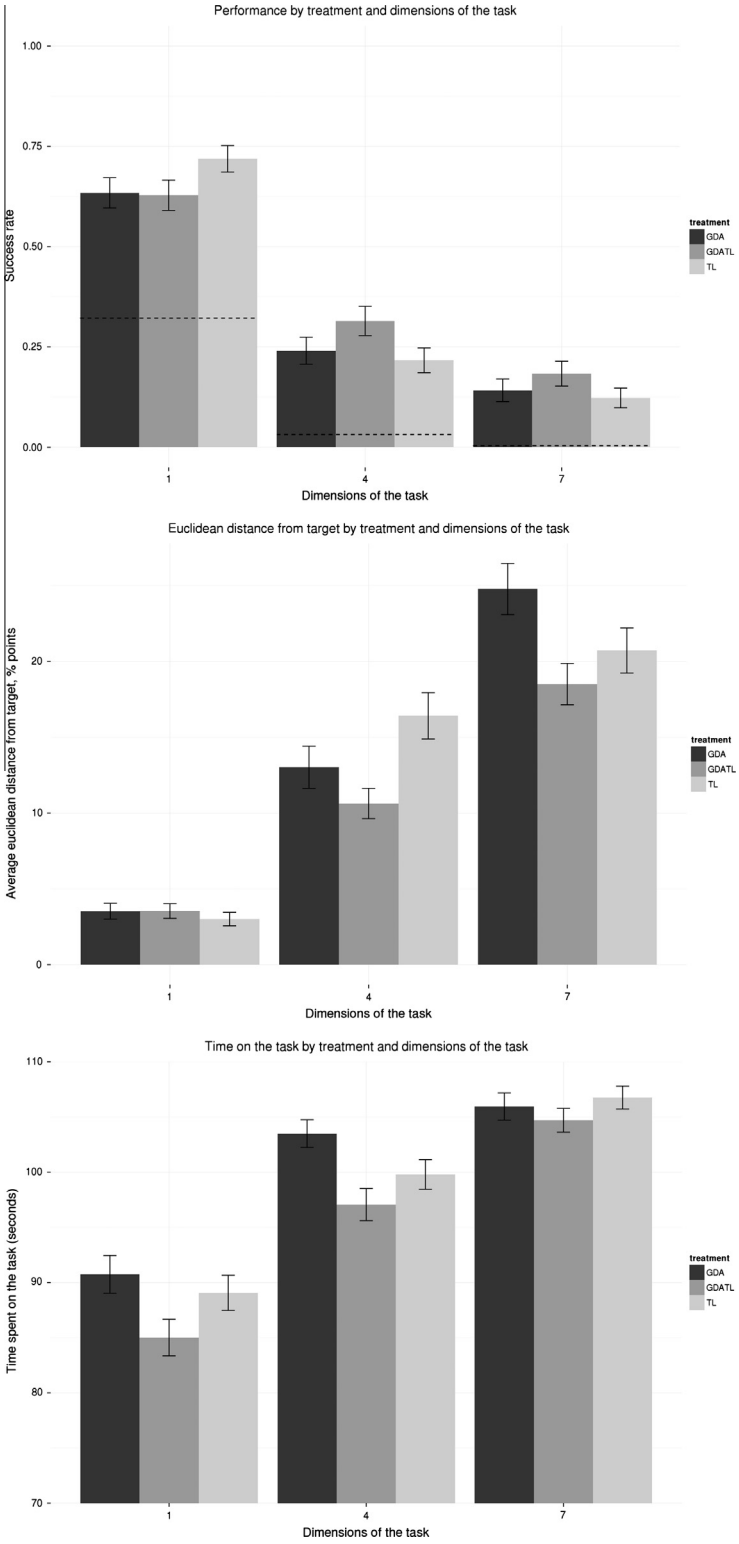


Fig. 6. Average % of correct tasks, by treatment and dimensions.

Table 8
Regression analysis of Experiment 2 – success rate and euclidean distance form target.

	Success rate random effect logit		Distance linear regression model	
Constant			−4.263	(−0.53)
TL	0.113	(1.44)	−5.446	(−1.90)
GDATL	0.112	(1.36)	−4.723	(−1.78)
4 dimensions	−0.246***	(−7.06)	7.459***	(4.54)
7 dimensions	−0.486***	(−10.99)	31.51***	(14.75)
Time	−0.0119	(−0.48)	5.045***	(3.72)
Interactions				
TL×4D	−0.0827	(−1.67)	2.262	(1.00)
TL×7D	−0.0509	(−0.82)	−2.345	(−0.93)
GDATL×4D	0.0207	(0.42)	−2.131	(−1.08)
GDATL×7D	−0.00302	(−0.05)	−2.692	(−1.13)
Performance in control tasks				
maths	−0.00598***	(−3.62)	0.145***	(4.74)
maths×TL	0.000844	(0.34)	−0.0220	(−0.46)
maths×GDATL	0.000860	(0.38)	0.0654	(0.80)
maths:time	0.0104	(0.40)	−1.366	(−1.23)
preferences	0.000865	(1.18)	0.0396	(1.50)
preferences×TL	−0.000915	(−0.88)	0.0671	(1.76)
preferences×GDATL	−0.000783	(−0.75)	0.0404	(1.23)
health	−0.000105	(−0.24)	0.0961***	(5.54)
Controls				
Female	−0.0834**	(−3.12)	0.846	(0.81)
Age	−0.00240*	(−2.01)	0.0466	(1.19)
Year Education	0.0145*	(2.25)	−0.416	(−1.81)
Income ('000€/month)	0.000171	(0.01)	−0.444	(−1.06)
BMI	0.00188	(0.73)	0.0268	(0.37)
Foodbudget (€)	−0.000189	(−0.82)	−0.00145	(−0.21)
Snacking	−0.00494	(−0.16)	1.702	(1.34)
n subjects	142		142	
N	1846		1802	

t statistics in parentheses.
For logit estimations marginal effects at the mean are reported.
$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

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 diets 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 favorable 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 behaviors. 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.

Appendix A. 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 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 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.

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

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 10
Reference thresholds to convert GDA into TL.

	Sugar (%)	Salt (%)	Fat (%)	Vitamin C	Fiber	Calcium
Green	<5.55	<5	<4.28	–	–	–
Red	>16.6	>25	>28.57	–	–	–

Appendix B. Complete statistical tests

Tables 11–13.

Table 11
Mann–Whitney Rank-Sum and χ^2 tests for Experiment 1 – INP students.

Variable	Dimensions	Test	Mean GDA	Mean GDATL	Mean TL	MW p-value	χ^2 p-value
Success rate	1	GDATL v GDA	0.95	0.94		–	0.732
		TL v GDA	0.95		0.93	–	0.657
		TL v GDATL		0.94	0.93	–	0.916
	4	GDATL v GDA	0.91	0.84		–	0.151
		TL v GDA	0.91		0.55	–	**0**
		TL v GDATL		0.84	0.55	–	**0**
	7	GDATL v GDA	0.86	0.78		–	0.151
		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**	–

In bold p-values that are under the 0.05 threshold.

Table 12
Mann–Whitney Rank-Sum and χ^2 tests for Experiment 1 – general population.

Variable	Dimensions	Test	Mean GDA	Mean GDATL	Mean TL	MW p-value	χ^2 p-value
Success rate	1	GDATL v GDA	0.71	0.74		–	0.753
		GDATL v TL		0.74	0.78	–	0.557
		GDA v TL	0.71		0.78	–	0.367
	4	GDATL v GDA	0.56	0.48		–	0.351
		GDATL v TL		0.48	0.35	–	0.15
		GDA v TL	0.56		0.35	–	**0.018**
	7	GDATL v GDA	0.41	0.29		–	0.139
		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**	–

In bold p-values that are under the 0.05 threshold.

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

Variable	Dimensions	Test	Mean GDA	Mean GDATL	Mean TL	MW p-value	χ^2 p-value
Success rate	1	GDA v GDATL	0.63	0.63		–	0.909
		TL v GDA	0.63		0.72	–	0.09
		TL v GDATL		0.63	0.72	–	0.07
	4	GDA v GDATL	0.24	0.31		–	0.137
		TL v GDA	0.24		0.22	–	0.596
		TL v GDATL		0.31	0.22	–	**0.04**
	7	GDA v GDATL	0.14	0.18		–	0.319
		TL v GDA	0.14		0.12	–	0.608
		TL v GDATL		0.18	0.12	–	0.121
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**	–

In bold p-values that are under the 0.05 threshold.

Appendix C. English translation of the original French instructions

Experiment 1

Welcome.
This experiment is ran by the GAEL laboratory of the University of Genoble and by INRA.

Privacy
In order to grant you anonymity during the session and the data analysis, we have given you a code.

Lab rules
All communication among the participants is forbidden, as well as all comment on how to behave in this experiment. Please stay focused on your PC screen and do not talk for the whole duration of the experiment. If you have any problem or question, please raise your hand and ask the experimenters.

Incentives
You have received an envelope that contains your show-up fee of 10 euro cash. This amount is yours to keep. Over and above this amount, and depending on your decisions, you will have the chance to earn extra money. This extra money will be paid to you in cash at the end of the session. You will get to know your payoff at the end of the experience. A summary of the decisions you took during the session and the associated payoffs will be shown to you.

Plan of the session
Before the start of the experimental

- you will be given instructions;
- you will be exposed to an example to test your understanding. This example will not be taken into account to compute your final payoff.

Your payoffs in this experiment will be a function of your choices.

Your task
We ask you to play the role of a nutritionist. You have been hired by a firm to compose daily menus for its employees. A daily menu is composed by:

- a compulsory daily basis (bread, butter, olive oil)
- breakfast, made up of
  - a drink
  - a solid food item
  - a fruit or juice
- lunch, made up of
  - entrée
  - main course
  - dessert
- Afternoon snack
- Dinner, made up of
  - entrée
  - meat or fish
  - side
  - dessert

For each of the items in the daily menu, you will have to choose one among 4 different options.

Example and exercise 1D
Your goal is to build the healthiest possible daily menus. Your work will be examined by an expert in nutrition. According to his recommendations, the total energy intake of the daily menu must lie between 1800 and 2200 kcal. Your daily menu will be considered healthy if it satisfies this condition.
To help you with your task, the energy content of each option will be displayed, as a percentage with respect to the recommended daily amounts. For instance, if a product contains 200 kcal, and since the recommended daily amount is of 2000Kcal, the product will be marked as providing 10% of the total energy intake of the daily menu.

Your task will be considered as fulfilled if the total energy intake of the daily menu lies between 90% and 110% of the recommended amounts. Please note: do not choose your preferred products. You will not buy nor consume the daily menu. Your goal is to build a nutritionally ideal daily menu for a third party.

The quantities of the products have been aligned to standard portions. A compulsory and common base of bread, oil and butter will be present in all daily menus.

To practice the task, you will first go through a learning exercise. This exercise is not relevant for your final payoff. You will learn your performance in the task as soon as you complete it.

[Subjects go through the 1D example screen]

Example and exercise 4D

The nutrition expert can set you more goals.

A daily menu will be considered healthy if, on top of meeting the energy constraint, it contains:

- less than 90 grams of sugar;
- less than 20 grams of saturated fat;
- less than 6 grams of salt.

[GDA only] To help you with your task, the energy content, as well as the content in sugar, saturated fat and salt, of each option will be displayed, as a percentage with respect to the recommended daily amounts.

[TL only] To help you with your task, the energy content of each option will be displayed, as a percentage with respect to the recommended daily amounts. Moreover, a color-code will also describe the sugar, saturated fat and salt contents of the products.

The color for the product will be

- GREEN if the product contains less than 5% of the recommended daily amount;
- AMBER if the percentage lies between 5% and 25%;
- RED if the percentage is higher than 25%.

[GDATL: displays both TL and GDA instructions, with small changes]

Your task will be considered as fulfilled if:

- the total energy intake of the daily menu lies between 90% and 110% of the recommended amounts;
- the total sugar intake of the daily menu is lower than 100%;
- the total saturated fat intake of the daily menu is lower than 100%;
- the total salt intake of the daily menu is lower than 100%.

Please note: Do not choose your preferred products. You will not buy nor consume the daily menu. Your goal is to build a nutritionally ideal daily menu for a third party. The quantities of the products have been aligned to standard portions. A compulsory and common base of bread, oil and butter will be present in all daily menus.

To practice the task, you will first go through a learning exercise. This exercise is not relevant for your final payoff. You will learn your performance in the task as soon as you complete it.

[Subjects go through the 4D example screen]

Example and exercise 7D

Finally, the nutrition expert can set you even more goals.

A daily menu will be considered healthy if, on top of meeting the energy constraint, it contains:

- less than 90 grams of sugar;
- less than 20 grams of saturated fat;
- less than 6 grams of salt.

But also:

- more than 80 milligrams of Vitamin C;
- more than 25 grams of fiber;
- more than 800 milligrams of calcium.

[GDA only] To help you with your task, the energy content, as well as the content in sugar, saturated fat and salt, vitamin C, fiber and calcium, of each option will be displayed, as a percentage with respect to the recommended daily amounts.

[TL only] To help you with your task, the energy content of each option will be displayed, as a percentage with respect to the recommended daily amounts. Moreover, a color-code will also describe the sugar, saturated fat and salt contents of the products. The color for the product will be

- GREEN if the product contains less than 5% of the recommended daily amount;
- AMBER if the percentage lies between 5% and 25%;
- RED if the percentage is higher than 25%.

Color-codes will also describe the amounts of Vitamin C, fiber and calcium present in each product.

Attention! The colors for vitamins, calcium and fiber are inverted with respect to the ones used for salt, sugar and saturated fat. This is due to the fact that vitamins, fiber and calcium are good for your health.

Hence, the color-codes will be

- GREEN if the product contains more than 25% of the recommended daily amount;
- AMBER if the percentage lies between 5% and 25%;
- RED if the percentage is lower than 5%.

[GDATL: displays both TL and GDA instructions, with small changes]

Your task will be considered as fulfilled if:

- the total energy intake of the daily menu lies between 90% and 110% of the recommended amounts;
- the total sugar intake of the daily menu is lower than 100%;
- the total saturated fat intake of the daily menu is lower than 100%;
- the total salt intake of the daily menu is lower than 100%;
- the total vitamin C intake of the daily menu is higher than 100%;
- the total fiber intake of the daily menu is higher than 100%;
- the total calcium intake of the daily menu is higher than 100%.

Please note: do not choose your preferred products. You will not buy nor consume the daily menu. Your goal is to build a nutritionally ideal daily menu for a third party. The quantities of the products have been aligned to standard portions. A compulsory and common base of bread, oil and butter will be present in all daily menus. To practice the task, you will first go through a learning exercise. This exercise is not relevant for your final payoff. You will learn your performance in the task as soon as you complete it.

[Subjects go through the 7D example screen]

Start of the experiment

The instructions and learning exercises are now over.

15 screens, similar to those of the learning exercises, will be displayed on your screens. For each successful completed task you will earn 1.50€. Along the 15 tasks you will not be informed about your results. You will be informed only at the end of the experiment.

Important! An objective is active on the screen if and only if the information needed to fulfill it is visible. For example, if only energy information is displayed, then your task will be considered successful if the daily menu you compose will have a total energy intake that lies between 90% and 110% of the recommended amounts. The other goals (sugar, salt, ...) will not be taken into account.

End of the experiment

Once you will be finished with the 15 tasks, you will

1. see a summary screen showing you your results in each of the 15 tasks;
2. be asked to fill in a short questionnaire;
3. once the questionnaire finished, you will see a summary of your gains in Euro;
4. we will then call you to be individually paid in the private, adjoining payment room.

Experiment 2

[Instructions for experiment 2 are very similar to instructions of experiment 1. Here we provide only the instructions for the extra image-only and mathematical tasks, and indicate the main changes in the rest of the instructions]

Preference task

Your goal is to compose your preferred daily menu. You have 5 min to complete this task. A timer will be visible on your screen. Once you have chosen your 11 preferred products, click on the "Submit" button. If you fail to do so before the time ends, your choices will not be recorded. This task will not be relevant for your final payoff.

Health task

We ask you to play the role of a nutritionist. You have been hired by a firm, and your task is to compose the healthiest possible daily menu for its employees.

As in the previous task, you will have for options to choose from for each of the 11 free courses of the daily menu. On the screen you will see the same products of the previous task. You have 5 min to complete this task. A timer will be visible on your screen. Once you have chosen your 11 preferred products, click on the "Submit" button. If you fail to do so before the time ends, your choices will not be recorded. This task will not be relevant for your final payoff.

Please note: in contrast with the previous task, here your taste and preferences do not have to be taken into account. Your task is to build the healthiest possible diet for a third party.

The quantities of the products have been aligned to standard portions. A compulsory and common base of bread, oil and butter will be present in all daily menus.

Main task

[In the main task instructions are identical to those of Experiment 1, but for three changes: subjects are reminded that they have 2 min for each task, that they have to submit in the given time or their choices will be lost; they are reminded that they have 15 s pause in between each task; and they will face only 9 screens, each of which can give rise to a payment of 2.5 Euro. The actual instructions for Experiment 2 are available upon request.]

Mathematical task

[1D] We ask you to make a computation exercise. You are going to see on your screen for rows of 4 numbers. Your task consists in choosing one number for each row, with the goal that the sum of the numbers chosen lies between 90 and 110. You have 2 min to fulfill the task. Do not forget to click on the "Submit" button. If you fail to do so before the time ends, your choices will not be recorded. If you successfully complete the task, you will gain an additional 2.5 Euro.

[4D] Let's make the task a bit more difficult. Now, each cell will contain four numbers. We will call A the first number of each cell, B the second, C the third, D the fourth. Your task consists in choosing a group of 4 numbers for each row, such that:

- the sum of all numbers A lies between 90 and 110;
- the sum of all numbers B is lower than 100;
- the sum of all numbers C is lower than 100;
- the sum of all numbers D is lower than 100.

You have 2 min to fulfill the task. Do not forget to click on the "Submit" button. If you fail to do so before the time ends, your choices will not be recorded. If you successfully complete the task, you will gain an additional 2.5 Euro. Please note that the task will be considered successful if and only if all criteria have been satisfied.

References

Borgmeier, I., & 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., & 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., & Lazaridis, P. (2011). *Nutritional labeling*.Oxford: Oxford University Press, pp. 520–545.

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., & 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., & 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. (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., & Brownell, K. D. (2013). The science on front-of-package food labels. *Public Health Nutrition, 16*, 430–439.

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., & 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*, 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., & 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, Synovate, West Malling, UK.

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., & Seidell, J. C. (2012). Methodological quality of front-of-pack labeling studies: A review plus identification of research challenges. *Nutrition Reviews, 70*, 709–720.