European Review of Agricultural Economics pp. 1–34
doi:10.1093/erae/jbz037

Nutritional and economic impact of five alternative front-of-pack nutritional labels: experimental evidence

Paolo Crosetto*, Anne Lacroix, Laurent Muller and Bernard Ruffieux

University Grenoble Alpes, INRA, CNRS, Grenoble INP, GAEL, Grenoble, France

Received May 2018; editorial decision May 2019; final version accepted May 2019

Review coordinated by Carl Johan Lagerkvist

Abstract

An incentivised laboratory framed field experiment with 691 subjects examined the impact of five front-of-pack labels (Multiple Traffic Lights; Reference Intakes; HealthStarRating; NutriScore and Système d'Etiquetage Nutritionnel Simplifié) on food shopping within a catalogue of 290 products. Using difference-in-difference, we estimate the between-label variability of within-subject changes in the shopping's Food and Standards Agency aggregated nutritional score. All labels improve the nutritional quality (−1.56 FSA points on average). NutriScore is the most effective (−2.65), followed by HealthStarRating (−1.86). Behaviourally, subjects react mostly to the extreme values of the labels and not to intermediate values. Nutritional gains are not correlated with higher expenditure.

Keywords: nutritional labels, laboratory experiment, front of pack

JEL classification: C93, Q18

1. Introduction

Consumers make hundreds of food choices every day. For each product, consumers can access a variety of information: price, size, ingredients, nutritional facts, origin of the products, environmental quality, organic and fair-trade certifications. This enormous amount of information competes for the limited time, attention and cognitive resources of consumers, giving rise to choice overload effects (Malhotra, 1982; Iyengar and Lepper, 2000; Kivetz and Simonson, 2000; Schwarz, 2004).

Confronted with these large search costs, consumers often rely on appearance (Imram, 1999), habit and simpler heuristics (Scheibehenne, Miesler and Todd, 2007; Schulte-Mecklenbeck *et al.*, 2013). Yet, everyday food choices

*Corresponding author: E-mail: paolo.crosetto@gmail.com

have potentially large impacts on long-term well-being, public health and treatment costs (Withrow and Alter, 2011).

Front-of-pack nutritional labels (FOPL) are a public policy tool conceived for a context of choice overload. They give less information than back-of-pack nutritional tables and ingredient lists, choosing focal, summarised and aggregated messages instead. FOPL are designed to be simple and to stand out among the mass of competing signals, and are more likely to be understood and used by consumers than detailed, information-rich tables (Bialkova and van Trijp, 2010; Becker *et al.*, 2015). Different FOPL formats have been shown to require different cognitive skills (Helfer and Shultz, 2014; Crosetto, Muller and Ruffieux, 2016a) and result in differentiated impacts on food choices (Hersey *et al.*, 2013; Vlaeminck, Jiang and Vranken, 2014; Ducrot *et al.*, 2016; Neal *et al.*, 2017).

The conceptual frameworks of Grunert, Wills and Fernández-Celemín (2010) and van Kleef and Dagevos (2015) describe exposure, integration, evaluation and decision as four steps necessary for nutrition labels to have any effect on consumers.

First, consumers must be aware of the labels. The physical characteristics of the label are the key determinants of attention to labels (Bialkova and van Trijp, 2010). For instance, pictorial elements are better recognised than words (Viswanathan *et al.*, 2009) and the use of colours improves not only consumers' awareness but also perceived healthfulness (Schuldt, 2013; Huang and Lu, 2016; Prevost *et al.*, 2017).

Second, once consumers are exposed to FOPL, their effect will be mediated by consumer understanding. Here again, the labelling format plays an important role as simpler labels seem to be better understood. Consumers identify healthier foods better with the colour-coded Traffic Lights than with the numeric Reference Intakes (Borgmeier and Westenhoefer, 2009; Kelly *et al.*, 2009; Balcombe, Fraser and Falco, 2010; Hawley *et al.*, 2013). Regarding the information focus, that is whether the label has multiple entries (nutrient approach) or is aggregated (product approach), simplicity still prevails although not as clearly as colour use. In cross-country studies where participants had to classify products, Egnell *et al.* (2018) found that the aggregated NutriScore performed better than multiple Traffic Lights and Reference Intakes but Feunekes *et al.* (2008) did not find any difference between 'seal of approval' and multiple Traffic Lights labels. Fundamental works in cognitive sciences give some insights on why colour-coded aggregate labels are cognitively the less effortful and the quickest to process (Muller and Prevost, 2016).

Third, based on their understanding, consumers may then use the label information to assess their food options. Numeric labels are by construction more informative than colour-coded labels and are *a priori* better equipped to provide consumers with the relevant information to compose a healthy diet. On the other hand, consumers may be reluctant to make the necessary cognitive efforts and may be more responsive to directive labels that allow the use of simpler heuristics. By abstracting away from exposure and preference, Crosetto, Muller and Ruffieux (2016a) show that Reference Intake labels outperform Traffic Lights in diet-building exercises but only when subjects were not time constrained.

Fourth, based on awareness, understanding and assessment, consumers eventually choose what to buy, depending on their preferences, budget constraints, shopping convenience and time available. This final step, the most important and the easiest to observe, is the focus of this paper. In a laboratory experiment, we examine the relative between-subjects effect of five labelling formats on within-subjects changes in food purchases in a context with high levels of exposure to, integration and evaluation of labels. Labels are presented and explained to all subjects so that the awareness, understanding and importance given to labels are magnified in the laboratory. While this might undermine the external validity of the absolute impact of labels on food purchase observed in the laboratory it allows us to focus on the relative effects of labels on purchase decision by giving the same level of exposure and understanding to all consumers across all labelling formats.

Several governments (Norden, 2010; FSA, 2013; Australia, 2013a) and the EU (Council of the European Union, 2011) have endorsed FOPL as a policy tool. Recently, France took the lead in Europe, legally committing to use a science-based procedure to choose which label to support. Article 14 of the 2016 French Halth Act[1] recommends food producers to adopt, on a voluntary basis, a uniform front-of-pack nutritional label design. Under the aegis of the health authorities, representatives of manufacturers, retailers, consumers and scientists were consulted and four candidate systems were selected: NutriCouleurs (a French version of the UK's Traffic Lights), NutriRepère (an updated version of the UK's Reference Intakes), NUtriMark (a French version of Australia and New Zealand's Health Star Ratings), the newly developed NutriScore (a five-colour mono-dimensional synthetic label) and SENS – (a frequency-based recommendation label). These candidate labels were then scientifically tested prior to adoption in a large-scale randomised controlled trial in French supermarkets (Ministère des Affaires Sociales et de la Santé and Fonds Français pour l' Alimentation et la Santé, 2017) and in the laboratory experiment described in this paper.

From a public policy perspective, the best FOPL is the one generating the healthiest diet changes at the lowest adjustment cost (see Rao *et al.*, 2013, for robust evidence that healthy food tends to be costlier). In this paper, we assess in a controlled laboratory experiment the nutritional and economic impact of the five different FOPL that were considered to be adopted in France. Six hundred and ninety-one subjects perform a real shopping task on a selected catalogue of 290 food products. We use a difference-in-difference design, whereby each subject, randomly assigned to one of six treatments (a benchmark plus one for each label), shop twice, once without and a second time, unannounced, with FOPL. By varying the label between subjects, we

1 Officially: *Loi 2016–41 du 26 Janvier 2016 sur la modernisation du système de santé français*, or Act about the French Health System (Touraine law).

can clearly identify, for each subject, its effect on nutritional quality and food expenditure when compared to the reference (no label) shopping basket.

We measure the nutritional quality of the shopping basket using the nutrient profiling model developed by the UK Food and Standard Agency (Rayner, Scarborough and Lobstein, 2009, henceforth, FSA score). The score is computed for each product by assigning negative points for salt, saturated fatty acids, calories and sugar, and positive points for fibre, fruit and vegetable content and proteins. Since the score is a scalar, it can be easily aggregated over baskets and compared across label treatments. We measure the food expenditure with two different indicators: the total expenditure on the food basket and the expenditure per unit nutritional improvement. The second indicator allows us to compare the efficiency of each label in generating nutritional change at low extra expenditure. The control provided by a laboratory experiment allows us to collect detailed behavioural, social and economic information on each subject. We complement this information with a label-specific qualitative questionnaire, in which we let subjects evaluate the labels' trustworthiness, appeal, ease of use and understanding.

We find that all labels but Reference Intake have a significant positive impact on the nutritional quality of the shopping. On average, labels improve the FSA score by 1.56 points, with respect to the baseline shopping basket with no labels. Colour-coded labels providing aggregate nutritional information generate nutritionally healthier choices than analytical labels using numbers. In particular, the NutriScore label (aggregate, colour) outperforms all others. It is followed by NutriMark (aggregate and analytic, black and white). We further find that the alternative approach of giving the consumer information about the suggested frequency of consumption (SENS label, aggregate and colour) performs worse than labels directly providing aggregate nutritional information.

Nutritional improvements are not associated with higher expenditure. When looking at total expenditure, NutriMark and NutriCouleurs result in overall lower expenditure – but they are the only two treatments in which the average size of the shopping basket decreased. When controlling for the nutritional gain and as a difference from the benchmark treatment, the additional cost per unit nutritional gain is not different from zero for any labels.

Behaviourally, results show that subjects further simplify the already simple messages conveyed by the label. For the three aggregate labels, subjects tend to focus on the extreme values, treating information on nuanced 5- or 10-level scales as binary (good–bad) or ternary (good–average–bad) information. These tendencies result in baskets that contain more products that are labelled 'green' or 'five star' and less products that are labelled 'red' or 'zero stars' but do not differ much for all the in-between categories. For analytic labels, subjects disproportionately focus on fat. Moreover, labels differ in their ability to crowd out other information, as proxied by the number of times subjects consulted the nutritional table and ingredient lists while shopping. NutriScore and Traffic Lights act as nearly perfect substitutes for detailed information tables. SENS, on the other hand, acts as a rather imperfect substitute, not being different from the benchmark with no label. The ability of labels to crowd out other information correlates with overall nutritional performance, and with the results of the qualitative questionnaire.

Subjects seem to comply more with labels that give clear, discrete, aggregate, colour-coded nutritional information. Labels that stand out and are cognitively easy to parse appear as more trustworthy and have a stronger nutritional impact than analytical labels that give a plenty of information but no focal, clear nutritional judgement. The cost of the adjustment is not different from zero for all labels. While the external validity of our laboratory experiment needs to be carefully assessed, our data suggest that FOPL are an efficient policy tool, allowing consumers to achieve considerable nutritional improvement at little cost.

2. Related literature

Several previous studies have measured the relative performance of different FOPL (see the review papers of Grunert and Wills, 2007; Vyth *et al*., 2012; Hersey *et al*., 2013; Volkova and Mhurchu, 2015; Cecchini and Warin, 2016). Most studies rely on experimental designs rather than observations on real-world data. This is because shopping data are usually difficult to organise, expensive, carry a lot of noise, suffer from severe selection biases and identification is usually hard to obtain. Still, some studies did use supermarket scanner data (Variyam, 2008; Boztuğ *et al*., 2015; Mørk *et al*., 2017). Studies follow by and large one of the four methods: declarative studies, randomised controlled trials, choice experiments and willingness-to-pay studies.

Surveys, questionnaires and general declarative studies set up hypothetical choice scenarios, and systematically vary the use of labels to identify effects. The subjects' task is usually to identify the healthier product among a set of two or three (Synovate, 2005; Borgmeier and Westenhoefer, 2009; Kelly *et al*., 2009). These studies allow researchers to assess the understanding and parsing of FOPLs. They do not allow to see how this understanding interacts with preferences; moreover, by relying on scenario choices, they suffer from hypothetical bias and do not generally take into account the role of prices in the choice. Recent surveys (Aschemann-Witzel *et al*., 2013) address the interaction of preferences and labels and extend the choice set up to 20 choices. They still retain a hypothetical nature, do not consider prices and usually focus on single-product categories.

Willingness-to-pay studies use incentivised elicitation methods (mainly experimental auctions and contingent valuation methods, Lusk and Shogren, 2007) to assess the perceived value of different food items, with and without labels. They indirectly assess, through the value that a subject assigns to a labelled product, the importance the subject attributes to the label's message. Usually run in the laboratory, they feature strong control on confounding factor but forfeit most of the external validity. In addition, these experiments usually involve a small number of products. Examples are in the context of

GMO content (Noussair, Robin and Ruffieux, 2004) and Thunström and Nordström (2015) focusing on healthy labels.

A multivariate and multi-attribute version of WTP studies is embodied by choice experiments. Choice experiments improve on surveys by giving the subjects dozens of choices, systematically varying the attributes of the products – price, quality and label – and analysing the data in a structural random utility model. They allow the researcher to estimate the marginal role of each attribute in a context in which prices are taken into account. On the other hand, choice experiments usually rely on hypothetical choices and on strong behavioural assumptions. Applications of the choice experiment method to different types of labelling come from Balcombe, Fraser and Falco (2010), who focus on Traffic Lights; Marchi *et al.* (2016), focusing on time preferences; Aprile, Caputo and Nayga (2012), dealing with quality labels; Loo *et al.* (2011), focusing on organic labels and Gracia, Loureiro and Nayga (2009), focusing on the value of nutritional tables.

Natural field experiments apply FOPL in grocery stores following a randomised procedure. Subjects might or might not be informed that they are part of an experiment, and their response to FOPLs is observed in a real setting and according to real preferences and budget constraints. These studies have a high degree of external validity. Data are extremely noisy and the experimenters do not have tight control on all possible confounds (promotions, shops opening hours). As result effects are usually small, as is the case of the recent French field experiment (Ministère des Affaires Sociales et de la Santé and Fonds Français pour l' Alimentation et la Santé, 2017). This might reflect the fact that effects truly are small, but that our ability to identify effects and tell apart signal from noise is lower in a field context. Examples of field experiments include Berning, Chouinard and McCluskey (2011) applying labels to popcorn in five stores in California, Neal *et al.* (2017) using a smartphone application to track consumer choices in an Australian supermarket and Vyth *et al.* (2010) studying choices in a Dutch supermarket.

In this paper, important parts of the methodology used are new in the literature. We examine *ex ante* the impact on food purchases of five competing nutritional labels that were each suggested by different food stakeholders in France (retailers, food industry and public authorities). The objective is to determine which labelling format will be implemented, on a voluntary basis, at the front of the package, with the ability of the label to change behaviour towards a healthier diet as the main selection criterion. To do so, we use a randomised controlled trial that allows us to observe real-purchasing choices from a representative sample over a large set of products, as in natural field experiments, but in a controlled laboratory setting in which we minimise noise, as in choice experiment or willingness-to-pay studies. Six hundred and ninety-one real consumers buy food from 290 food items in all food categories. They shop twice, without and then with nutritional labels (only one randomly drawn basket is incentivised). This within-subject structure allows us to control for individual differences and, hence, increase statistical power.

While other studies have opted to run randomised trials in laboratory settings (Borgmeier and Westenhoefer, 2009; Roberto *et al.*, 2012a; Spanos, Kenda and Vartanian, 2015), in real settings (Crockett *et al.*, 2014; Gaigi *et al.*, 2015) and online (Roberto *et al.*, 2012b; Waterlander *et al.*, 2013; Watson *et al.*, 2014; Wong *et al.*, 2014; Vasiljevic, Pechey and Marteau, 2015; Ducrot *et al.*, 2016), we are, to the best of our knowledge, the only study that ensures ceteris paribus conditions on such a large scale. We are thus able to provide clean, incentivised laboratory evidence on the relative effect of different FOPLs on food purchases, abstracting away from exposure and understanding. Our design allows us both to assess the nutritional quality of the whole shopping with and without competing labels, and to observe in detail the behavioural sources of this nutritional change. Finally, we are the first to test a frequency-based label (SENS) rather than the product- or nutrient-based labels tested so far.

3. Method

Subjects were asked to shop for 2 days for their household, following their usual shopping habits. A paper catalogue of 290 products, divided in 39 categories, was distributed to each subject. The catalogue displayed for each product a full colour picture, price, weight (*g*) or volume (*l*), price per kg or per litre and a bar code. Prices corresponded to those recorded before the experimental campaign in a local supermarket. Using a bar-code reader, subjects could display on their screen the product of their choice in a custom online e-shopping environment. They could thereby access, for each product, a list of ingredients (with outlined allergens) and a nutritional table.

Subjects were asked to shop twice; once with a benchmark catalogue without nutritional labels and a second, previously unannounced time with a new catalogue, strictly identical to the previous one but in which nutrition labels were added. In the second catalogue, all products that are legally subject to labelling were labelled, i.e. all products except fresh fruits and vegetables, fresh packaged meat and eggs.

We ran six between-subjects treatments: one for each of five nutritional labels and a benchmark neutral treatment in which no label was added and subjects shopped twice with the same catalogue.

The experiment was incentive compatible. Subjects were informed that they would have to buy a subset of one of their two shopping baskets. The payoff-relevant basket would be randomly and publicly drawn at the end of the experiment. The subset of products actually put up for sale was unknown *ex ante*. In a separate room, we had stored about a quarter of all the catalogue products. The intersection of the items selected by the subjects and what we had in store was then sold, at catalogue prices, to the subjects at the end of the session.

4. Measures

Our experimental design (already put to use in its general form in Muller and Ruffieux, 2012; Muller *et al.*, 2017) allows us to measure behaviour twice. The first, unlabelled basket allows us to set a benchmark for the shopping behaviour of each subject. The second basket allows us to assess, within subjects, changes with respect to the baseline basket. The comparison of individual changes across treatments allows us to cleanly assess, by exploiting differences in differences, the effect of labels by controlling for the heterogeneity of individual preferences.

The main measure of interest is the aggregation per treatment of the individual change in the nutritional score between basket 2 (labelled) and basket 1 (unlabelled). We adopt as our nutritional measure the Nutrient Profiling Model developed by the UK Food and Standard Agency (Rayner, Scarborough and Lobstein, 2009, henceforth FSA score). This score is computed for each product by assigning negative points for salt, saturated fatty acids, calories and sugar, and positive points for fibre, fruit and vegetable content and proteins. The score ranges from −15 to 35, with lower numbers indicating better overall nutritional quality.

We compute the aggregate nutritional score for each shopping basket, by adding the FSA score of each item and normalising by 100 kcal. That is, for each subject i, for each basket $j \in [1,2]$, we compute

$$\mathrm{FSA}_{ij} = \frac{\sum_p \mathrm{kcal}_{pij} \times \mathrm{FSA}_{pij}}{\sum_p \mathrm{kcal}_{pij}},$$

in which the index pij denotes each product p in caddy j for subject i.

This measure gives us a single, continuous variable to assess the nutritional quality of the whole shopping basket. Nonetheless, it relies on two assumptions: that the FSA score correctly assesses the nutritional value of a shopping basket and that a normalisation by calories is not distortive. The FSA is widely used in epidemiological studies (Julia *et al.*, 2014, 2015b; Ducrot *et al.*, 2015) and its value as a nutritional score is well documented. Normalisation by energy content has the advantage of delivering results that are easily interpreted as 1 day of consumption for an average adult, and to anchor the results to the basic measure of human energetic needs. Nonetheless, studies vary as per the normalisation used. For robustness, we ran our analysis also using SAIN/LIM, an alternative nutritional system (Tharrey *et al.*, 2017) and normalising by weight.

We estimate the treatment effects using a difference-in-difference regression analysis. We let the treatment and basket variables interact, and add a series of subject-specific variables to control for the socio-demographics of the sample. Our main estimation uses data from both baskets and all six treatments and takes the following form:

$$\mathrm{FSA}_{it} = \beta_0 + \beta_1 \mathrm{Basket2} + \delta \mathrm{Treatment} \times \mathrm{Basket} + \gamma \mathrm{Controls},$$

in which the benchmark treatment with no labels serves as a reference category, β_0 identifies the average nutritional value of the first basket, β_1 the impact of the benchmark treatment, δ is a vector of 5 × 2 coefficients identifying the difference with respect to the first benchmark basket of the first and the second basket for each of the five labels (NS, NR, NC, NM and SENS) and γ is a vector of subject-specific coefficients.

We supplement the analysis of nutritional impact with economic considerations. First, in line with public policy concerns for lower income households, we separately run regressions for three different standard-of-living groups. Our recruitment was stratified by income. To properly account for the different size and composition of households, as done in food research by Nordström and Thunström (2011a, 2011b), we compute three standard-of-living classes based on the income by consumption unit. While elsewhere in the literature, household income is normalised by the number of adult earners in the household, we use a different normalisation that gives higher importance to the number of children. We use the official definition of 'standard of living' by consumption units of a household provided by the French Statistical Bureau.[2] Our subjects can, hence, be divided into three standard of living groups of roughly equal size, characterised by living standards <100/month, 1,000–1,500/month and >1,500/month (see Table A.9 in Appendix A in supplementary data at *ERAE* online for the breakdown by treatment).

Second, we analyse the change in total expenditure associated with the nutritional gains, to see whether the nutritional improvement comes at the cost of higher average expenditure. As a first approximation, we compute the total expenditure by basket as the simple sum of all the prices of all the products bought. We then analyse the relationship between the nutritional gain (in scoreFSA) and the economic cost (in extra expenditure), exploiting our difference-in-difference design once more.

In addition to these analytical measures based on the FSA score, we compute a number of behavioural and qualitative measures to identify the factors driving the main nutritional and economic effects. In particular, we track the number of items in each basket, the number of product entries and exits from basket 1 to basket 2, the number of clicks on the ingredient lists and nutritional tables. We complement this with label-specific behavioural variables keeping track of the qualitative change between baskets. Finally, we asked subjects to fill in two questionnaires, one aimed at assessing their understanding of the labels and another including socio-demographic characteristics.

2 INSEE, see the definition of 'standard of living' (niveau de vie) as income divided by consumption units, details https://www.insee.fr/en/metadonnees/definition/c1890here. We divide the total household income declared by our subjects in the final questionnaire by the number of consumption units in the household. Consumption units are 1 for the first adult in the family, 0.5 for any additional person older than 14 and 0.3 for all persons under 14.

5. Experimental details

5.1. Products

Subjects could shop in a printed colour catalogue of 290 food items grouped into 39 categories.[3] While not being perfectly representative of shopping in a real supermarket, where thousands of products are available, the catalogue covered all the needs of an average household and included fresh, canned, packaged and frozen food, fruits and vegetables, snacks and mixed prepared dishes. The overall coherence and representativeness of the catalogue were, respectively, verified by nutritionists and using actual consumption data from the Kantar World Panel. All products were currently on sale in local supermarkets at the time of the experiment. The prices used in the catalogue were collected in a specific supermarket the week before the start of the experimental campaign, in October 2016. The catalogue displayed a picture of the front-of-pack of the product, its name, price, weight or volume and the price per kilogram (or litre). The labelled version of the catalogues featured a label below the product picture. The labels were more prominent in the catalogue than one would usually find in a supermarket, and more similar to the kind of information that could be available on an online platform. This was intentional. In real supermarkets, it is not possible to control whether consumers have paid attention to the label – hence, the effect of the label on behaviour is mediated by the exposure and understanding of the label by the consumer. In our experiment, we wanted to fully capture the labels' effects on purchases when understanding and exposure concerns are out of the picture.

5.2. Subjects

The study involved 691 subjects over 42 sessions (7 per treatment). Subjects were recruited among the general population of the Grenoble metropolitan area by a professional recruitment agency. Grenoble and its suburbs, located in the Alps in south-eastern France, has about 400,000 inhabitants. Recruited subjects were in charge of the grocery shopping for their household and regular supermarket customers. The sample was stratified by household disposable income: one-third of subjects with less than 2,000 Euros per month, one-third between 2,000 and 3,000, and another third with more than 3,000. Following standard procedures of randomised controlled trials, individuals from each income group were randomly allocated to either one of the five label treatments or the benchmark treatment. Summary statistics for our sample are provided in Table A.9 (in Appendix A in supplementary data at *ERAE* online). Overall, one in five subjects was a man, the age and income structure loosely reflected those of the Grenoble Metro Area. Randomisation

3 For screenshots of the catalogue, see 11. The full catalogue for the benchmark treatment is available at https://www.dropbox.com/s/jcrd11rq2cy5y3x/2016-11-15_label4_neutre.pdf?dl = 0

checks did not reveal significant differences across treatments, except for the distribution of professional status.

5.3. Tested labels

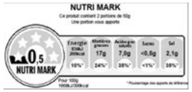
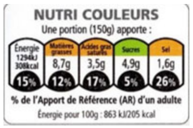
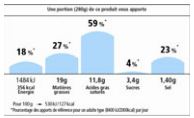
We tested five nutritional labels (see Table 1 for an overview). The choice of labels to be tested was determined in the French Health Bill, with different labels being proposed by different stakeholders. Omitting binary recommendation labels as the keyhole (Norden, 2010) or the Heart Foundation Tick (Australia, 1989), these labels span the whole space of those in use today in different countries, and follow one or both of the main approaches in nutritional labels: providing detailed analytic information about a set of key nutrients, or displaying an aggregate assessment of the food item. The labels use colours, numbers and letters to convey their meaning; they rely on absolute, relative or frequency assessment and they are based upon different algorithms to translate bare nutritional values into marks and ranks.

NutriScore (NS) was developed and validated by an independent research team (Equipe de Recherche en Epid´emiologie Nutritionnelle Julia and Hercberg, 2017). NS is based on the nutrient-profiling system developed by the UK Food Standards Agency (Rayner, Scarborough and Lobstein, 2009, FSA score) later adapted to the French context Haut Conseil de la Santé Publique (2015). Its final graphical format was developed in a dedicated study (Nugier, Serry and Thanh, 2016). NS is an aggregate, colour-coded label, similar to the energy efficiency labels used in the home appliance sector, which gives coarse but salient information, in the form of a letter (from A to E) colour-coded from green to red. Studies have shown it to be effective in evaluating the healthiness of foods (Ducrot *et al*., 2015) and to impact purchasing intentions in an online supermarket (Ducrot *et al*., 2016; Julia *et al.*, 2016), in a controlled laboratory setting (Crosetto, Muller and Ruffieux, 2016b). Also as a consequence of the results of the present study, NutriScore was elected as the official French front-of-pack labelling system in October, 2017 (République, 2017).

NutriMark (NM) was adapted by the French retailer Leclerc based on the Australian Government Health Star Rating system (Australia, 2013a, 2013b; Cooper, Pelly and Lowe, 2017). NM displays both aggregate information, in the form of stars, ranging from 0.5 to 5 and nutrient-specific information in the form used by Reference Intake labels. NM was tested on Leclerc's online shopping platform in the autumn of 2016, involving 3,000 Leclerc brand products at 84 collection points in France. Results show a slight improvement in the average nutritional quality of purchases, but only for the middle-class and the under-30s.

NutriCouleur (NC) was developed under the original name Multiple Traffic Light system by the UK Food and Standard Agency (FSA, 2013). Its introduction in France was supported by Nestlé. NC presents analytical information for energy and four key nutrients (fat, saturated fatty acids, sugar and

Table 1. Nutritional labels tested in the experiment

Label	Name	A.K.A.	Information	Colour	Base	Reference	Use
	NutriScore (NS)	5Couleurs	Aggregate	Yes	FSA score	Julia *et al.* (2014)	(since 2017) France
	NutriMark (NM)	Health Star Rating	Aggregate and analytic	No	HSR score	Australia (2013a,b)	Australia, New Zealand, Leclerc retailer
	NutriCouleur (NC)	Multiple Traffic Lights	Analytic	Yes	nut.values	FSA (2013)	United Kingdom, Ireland, Spain
	NutriRepère (NR)	Guideline Daily Amount, Reference Intake	Analytic	No	nut.values	Rayner, Scarborough and Williams (2004)	EU industry standard, US facts upfront, …
	SENS	–	Aggregate	Yes	LIM score	Tharrey *et al.* (2017)	–

salt) in three different ways: as a percentage contribution to the daily reference intake, in absolute amounts per serving (in grams) and in colour on three levels (red, amber and green). Several studies have investigated the impact of Multiple Traffic Lights on choices (Kelly *et al.*, 2009; Bialkova and van Trijp, 2010; Aschemann-Witzel *et al.*, 2013; Julia *et al.*, 2016; Crosetto, Muller and Ruffieux, 2016a), showing overall greater effectiveness with respect to labels that do not use colours.

NutriRepère (NR) was proposed by the ANIA, the organisation representing the interests of the French food industry and grouping 500 food producers in France. It is based on the Reference Intake label (RI).[4] RI has been adopted by the Australian food and beverage industry in 2006 (as the 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). NR presents analytical nutritional information for energy and four key nutrients (fat, saturated fatty acids, sugar and salt). This information is displayed in three different ways: in percentage, absolute value and by means of light blue bars.

Système d'Etiquetage Nutritionnel Simplifiè (SENS) was developed by nutritionist Nicole Darmon and her group (Tharrey *et al.*, 2017) based on the previously established SAIN/LIM nutritional profiling (Darmon *et al.*, 2009). SENS was backed by the organisation of French distributors, grouping 50 supermarket chains. SENS presents nutritional information as a colour-coded recommendation of eating frequency, using a red–blue–orange–purple palette. That is, after assessing the food nutritional quality, SENS translates this index into a frequency, and tells the consumer if a food item might be eaten anytime, often, from time to time or rarely.

Crucially, none of the labels tested was in use at the moment of the experiment exactly in the format tested. Two labels, NutriRepère and NutriCouleurs, were similar enough to Guideline Daily Amounts and Traffic Lights to be possibly familiar to our subjects, but both were in a visually changed format, and TL had seen very little adoption in France, mainly limited to products imported form the UK. Subjects had no previous exposure to SENS, NutriScore nor NutriMark.

Overall, the labels vary across three dimensions that have been shown to be crucial by previous research (Drichoutis *et al.*, 2008; Shogren, 2011; Muller and Ruffieux, 2012; Muller and Prevost, 2016; Crosetto, Muller and Ruffieux, 2016a): (i) whether the information is aggregated (NS, SENS, in part NM) or analytical (NR, NC, in part NL); (ii) whether the label uses colour (NR, NC, SENS) or not (NM, NR) and (iii) whether the label focuses on nutritional content (NS, NM, NC, NR) or on consumption frequency (SENS). More generally, labels trade-off salience with detail, with NS and

4 Reference intakes have been previously known as Guideline Daily Amount (Rayner, Scarborough and Williams, 2004). Although the principles behind GDA and RI 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.

SENS choosing to deliver a salient but coarse message with no readily available reference to nutritional values, and NR and NC choosing to give more detailed information in a less salient package. NM includes both approaches in a combined design, but forfeits the use of colour.

5.4. Laboratory procedures

Subjects were invited for sessions that they knew would last approximately 1.5 hours and received 32 as a show-up fee. Their task was to shop to cater to the needs of their family for 2 days but were otherwise not directed in their choices. This frame was given to reach some uniformity in the task that the subjects faced, but it was not enforced. If subjects asked what '2 days' worth of consumption' meant, they were told that each must decide according to their household's taste and needs. Subjects were allowed to shop more, or less, or not at all.

Instructions were then shown on each subject's screen, as well as projected overhead, and read aloud by the experimenter.[5] Questions were asked and answered publicly all along the reading of the instructions. The English translation of the original French instructions is available in 10.

After all instructions had been given, including a screenshot-based demonstration of the software interface,[6] subjects were asked to do their shopping. They did so in individual cubicles, from which they could not see the other subjects' screens. Subjects were not given a time limit, and each could shop at his or her own rhythm. When all subjects had finished their shopping, they were asked to shop again in an unannounced second shopping exercise.

For all treatments but the Benchmark, after the first shopping and before the second one, subjects were given a one-page explanation of the nutritional label that they would face in the second shopping (see 10). These explanations were based on the flyers distributed in the 60 supermarkets of a natural field experiment comparing the same labels performed alongside the current experiment in the Fall of 2016 (Ministère des Affaires Sociales et de la Santé and Fonds Français pour l' Alimentation et la Santé, 2017). Subjects received a paper copy of these one-page explanatory sheets; the same was projected on a screen and its contents read aloud. Any question was addressed before proceeding with the second shopping period.

After each subject submitted his or her choices for the second shopping basket, one of the two was randomly selected as binding using a physical urn for increased transparency. Then subjects were exposed to a socio-demographic questionnaire and, for all treatments but the Baseline, to a qualitative survey asking them to rate over several dimensions the nutritional label they had been exposed to. Subjects left then the room individually, and

5 Due to the size of the experimental campaign, four different experimenters were involved. Results are robust to controlling for the identity of the experimenter.

6 The source code for the interface, written in PHP, is available upon request.

bought (a subset of) the items they had chosen in the binding basket in a separate room.

6. Results

6.1. Nutritional results

6.1.1. The benchmark treatment

In the benchmark treatment, subjects shopped twice under the same conditions. This treatment allows us to assess the degree of variability of diets across repetitions. Indeed, a repetition *ceteris paribus* does not necessarily have to generate the same choices. First, as basket 1 was not recalled on the screen when basket 2 was created, subjects had to reconstruct it from memory. Second, subjects might wish to change some items because of a taste for variety. Third, a basket does not reflect all eating habits but represents only a limited sample so it is natural to expect a second sample to differ. Finally, subjects could find repetition boring and make changes to escape boredom. Although we expect variability in all treatments, we do not expect any systematic nutritional effect in the benchmark treatment.

This is indeed what we observe. The average FSA score for basket 1 is 5.22 (SD 3.01) and 5.34 (2.94) for basket 2. The difference is not significant (Wilcoxon signed-rank test, p-value 0.77). Figure 2 plots for each subject the absolute difference in FSA score from basket 2 to basket 1 (negative changes mean healthier diets), for all treatments, and overlays the points with a boxplot. The distribution for the Benchmark treatment (light grey, far right) is symmetric and centred around zero. Despite a large variability in nutritional quality between baskets 1 and 2 within the benchmark treatment, there is no net effect on overall nutritional quality, as individual deteriorations and improvements cancel each other out.

6.1.2. The label treatments

The nutritional quality of basket 2 improved across the board with respect to basket 1 in all treatments with a label. Averaging over all label treatments, the mean effect was of −1.56 points (Wilcoxon signed-rank test, p-value < 0.001). Individually, all treatments but NutriRepère showed a significant effect at 5 per cent (WSRT, all p-values < 0.02), while NR was just short of 5 per cent (WSRT, $p = 0.0579$). A graphical representation of this main result can be seen in Figure 1, in which error bars represent 95 per cent confidence intervals. Mean treatment effects (absolute difference in FSA score between basket 2 and basket 1) are given along the diagonal in Table 2. The table also shows the results of pairwise comparing any two treatments. The upper triangle of the table gives the result of Wilcoxon rank-sum tests, the lower diagonal of t-tests. Results are robust to the tests used. Among the label treatments, NutriScore performs better than any other label. NutriMark comes in second, outperforming all other remaining labels but NutriCouleur. SENS and NutriRepère are not statistically different from each other and come last.

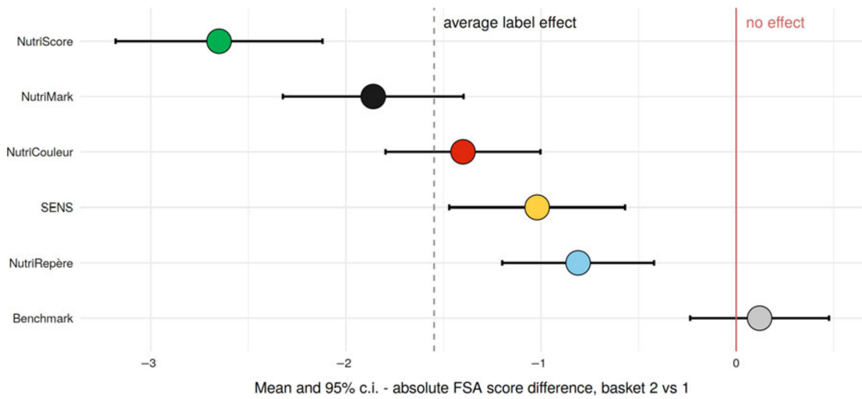


Figure 1. Average absolute FSA score change, basket 2 vs. basket 1, by treatment.

Table 2. Average treatment effects (FSA score) and *p*-values from *t*-tests and Wilcoxon rank-sum tests

	Benchmark	NS	NM	NC	SENS	NR
Benchmark	0.12	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**0.001**
NS	**<0.001**	–2.65	**0.034**	**0.001**	**<0.001**	**<0.001**
NM	**<0.001**	**0.048**	–1.86	0.26	**0.014**	**<0.001**
NC	**<0.001**	**0.001**	0.394	–1.40	0.142	**0.006**
SENS	**0.002**	**<0.001**	**0.041**	0.459	–0.81	0.222
NR	**0.019**	**<0.001**	**0.006**	0.236	0.492	–1.02

Mean treatment effect (FSA score absolute difference) on the diagonal (greyed).
p-values from two-tailed *t*-tests (lower triangle) and from Wilcoxon rank-sum tests (upper triangle).
Significant tests in bold.

As noted for the benchmark treatment, there is a natural between-basket variation in the nutritional quality of the shopping. In the labelled treatments, though, there is a switch to healthier products that are directly imputable to the label. The distribution of FSA score change by treatment (Figure 2) shows that in every treatment there is a share of subjects that worsened the overall nutritional score of their shopping (difference >0). This is, as in the benchmark treatment, an effect of the natural variability in the subjects' shopping. None of the labels reduces the variance of the FSA score changes (all changes in variance are positive, but none exceeds 5.6 per cent). Hence, no label reduces the intrinsic variability of diets. Rather, the whole distribution moves down (towards healthier scores).

The descriptive results are confirmed by our difference-in-difference estimation, as shown in Table 3, first column. In the regression, NR and SENS

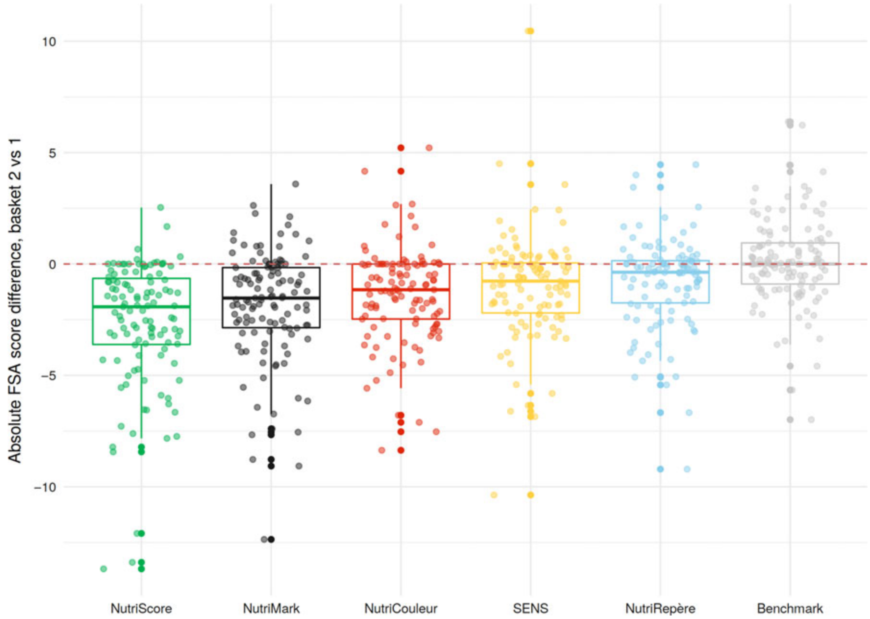


Figure 2. Distributions and boxplots of FSA score change, basket 2 vs. 1, by treatment.

fail to reach significance, but NS, NM and NC all have a significant impact. The results are robust to controlling for demographics (see Table B.11 in Appendix B, column 1 in supplementary data at *ERAE* online). Summarising, even if all labels induce nutritional change, we see a group of labels that consistently show significant results across different methods (NS, NM and NC), and two labels (NR and SENS) that show smaller, not significant results.

The impact of nutritional policies on households characterised by a low standard of living is usually the key to their success, since nutrition-related problems are often correlated with low income (Drewnowski and Specter, 2004; Drewnowski and Darmon, 2005; Drewnowski and Eichelsdoerfer, 2010).

The nutritional impact of the labels is summarised in Table 3, right panels, where our difference in difference specification is run separately by the standard-of-living group. The ranking of the point estimates of label effects stays constant, but NutriCouleur is not significant and for the high living standard (LS) group only NutriScore stays significant. This loss of significance is partly due to the lower number of subjects. Nonetheless, the high living standard group is also characterised by higher nutritional quality of basket 1 (lower intercept) and the coefficient for NutriMark is remarkably lower than for other groups, possibly indicating a different impact of NutriMark for the high living standard group.

Table 3. Difference in difference treatment effect estimations, overall and by income class

	Full sample	Low LS	Middle LS	High LS
Intercept	5.225***	5.532***	5.371***	4.552***
	(0.305)	(0.523)	(0.485)	(0.589)
Basket 2	0.115	0.026	0.259	0.001
	(0.431)	(0.740)	(0.686)	(0.833)
NutriScore	−0.487	−0.676	0.158	−0.490
	(0.437)	(0.764)	(0.770)	(0.760)
NutriMark	−0.054	−0.207	0.125	−0.017
	(0.434)	(0.723)	(0.723)	(0.819)
NutriCouleur	−0.846	−1.264	−0.132	−0.913
	(0.437)	(0.803)	(0.723)	(0.767)
NutriRepère	−0.430	−1.383	0.282	0.147
	(0.444)	(0.749)	(0.763)	(0.802)
SENS	−0.656	−1.316	−0.255	−0.085
	(0.441)	(0.776)	(0.784)	(0.760)
NutriScore × Basket 2	−2.766***	−2.646*	−2.914**	−2.675*
	(0.619)	(1.081)	(1.088)	(1.075)
NutriMark × Basket 2	−1.974**	−2.086*	−2.442*	−1.155
	(0.613)	(1.023)	(1.022)	(1.159)
NutriCouleur × Basket 2	−1.513*	−1.295	−1.549	−1.595
	(0.619)	(1.135)	(1.022)	(1.085)
SENS × Basket 2	−1.140	−0.900	−1.167	−1.224
	(0.624)	(1.097)	(1.109)	(1.075)
NutriRepère × Basket 2	−0.924	−0.724	−0.797	−1.199
	(0.627)	(1.059)	(1.079)	(1.135)
N	1,382	462	460	460
R^2	0.078	0.088	0.067	0.097
logLik	−3,632.179	−1,213.428	−1,213.813	−1,185.175
AIC	7,290.357	2,452.855	2,453.627	2,396.351

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$.

6.1.3. Robustness checks

Results might depend on the nutritional indicator or normalisation used for the analysis. We, hence, run robustness checks for both the descriptive and regression analyses.

The FSA score may favour some labels over others. In particular, for aggregate labels, NS and NM are directly built on the FSA score, while SENS is built on the LIM nutritional score (Darmon *et al.*, 2009). Insofar as the LIM and the FSA are not perfectly correlated, adoption of the FSA score as the determining indicator could adversely affect the performance of SENS relatively to NS and NM. For analytical labels, NC and NR give nutrient contents and are, therefore, dependent on the weightings made by the FSA score. We run robustness checks using as independent variable the LIM index for 100 kcal, and four individual nutrients: salt, saturated fatty acids, sugar and fat.

The normalisation used to aggregate the shopping basket might also prove distortive. Labels might correlate differently to quantities than to calories and, hence, reporting the shopping to energy intakes rather than weight could be a less than innocent exercise. We, hence, run the FSA score analysis normalising by weight (100 g) rather than calories.

Results are summarised in Table 4 (descriptive statistics and tests) and in Table B.10 (in Appendix B in supplementary data at *ERAE* online) (difference-in-difference regressions). Results are in general not impacted by the change in indicator or normalisation: NS, NM and NC consistently show significant results, while SENS and NR less so. The use of LIM instead of the FSA score does not change the results or the ranking of labels. Paradoxically, SENS performs even worse with LIM than with the FSA score. Labels do not induce significant changes in salt or sugar content. Fat and saturated fatty acids contents are instead impacted. As in the FSA score analysis, NS, NM and NC show significant impacts. NC performs particularly well, probably because of its analytical nature, which gives the consumer interested in a particular nutrient salient detailed information. Normalising by weight does not impact the ranking of labels either, excluding the fact that for some indicators, SENS underperforms NR.

6.2. Economic results

Nutritional gains might come at an economic cost, which could undermine the long-term impact of labels, as households would face a trade-off between increasing the nutritional value and facing their budget constraints. We start

Table 4. Robustness checks.

	Benchmark	NS	NM	NC	SENS	NR
LIM	0.15 (3.34)	−3.94*** (4.91)	−2.57*** (4.22)	−2.31** (3.9)	−1.39† (4.29)	−1.44† (3.68)
Fat	0.37 (12.83)	−6.85** (11.57)	−6.49** (12.69)	−7.25** (12.46)	−2.74† (15.13)	−3.65 (12.75)
Saturated fatty acids	−0.01 (9.33)	−5.96*** (9.92)	−5.17*** (10.66)	−5.3*** (8.81)	−0.64 (12.43)	−3.11* (10.52)
Sugar	0.23 (23.49)	−3.13 (18.97)	−1.75 (19.97)	−0.9 (22.02)	−2.84 (14.06)	−3.72 (18.03)
Salt	0.02 (0.67)	0.03 (0.71)	−0.15 (0.59)	−0.14 (0.61)	−0.01 (0.61)	−0.03 (0.71)
Normalised by weight	0.02 (1.04)	−1.21*** (1.45)	−0.85*** (1.24)	−0.52* (1.09)	−0.46* (0.98)	−0.54* (1.2)

Mean (standard deviation) absolute change in indicator, by treatment.
Wilcoxon signed-rank tests, significance thresholds: p-values: †<0.10; *<0.05; **<0.01; ***<0.001.

by considering the total expenditure on a basket. Subjects spend around 30 for each basket (Table 5, left). To test the difference in the total expenditure on a basket across shopping baskets by treatment, we can exploit the within-subject variation across baskets, and run a Wilcoxon signed-rank test (WSRT, Table 5, column 5). The total expenditure on a basket stayed the same for the Benchmark, NS, NR and SENS treatments and decreased for NM and NC.

Overall, then, the labels have little effect on total expenditure, except for NM and NC, that manage in both increasing the nutritional quality *and* decreasing the total expenditure on the basket. It might be noted that these two treatments are the only ones in which a significant (albeit small) decrease in the total number of products bought is found (see Section 6.3).

Our data allow us to directly test the relationship between the nutritional gain and the change in overall expenditure, across treatments. We can cleanly compare treatments using a difference in difference approach. Table 6 reports the results of the estimation:

$$\Delta \text{expenditure}_{\text{i}} = \beta_0 + \gamma(-\Delta \text{FSA}_{\text{i}}) \times \text{treatment},$$

in which the between-baskets change in overall expenditure, Δexpenditure, is regressed on the interaction of treatment and the between-baskets change in FSA score, ΔFSA. The Benchmark treatment acts as baseline and, to ease interpretation, we inverted the sign of ΔFSA so that higher values imply better nutritional content. The coefficients can be directly interpreted as extra Euro spent to obtain a one-point nutritional improvement.

None of the interaction effects is significant, meaning that despite the significant point estimates shown above for NM and NC, no treatment shows a significant difference in the reduction of expenditure for a unitary nutritional gain over and above what we observe in the Benchmark treatment. This is evidence of the fact that subjects reacted to the labels by performing *substitutions* within their budget constraint, adding healthier products at the expense of unhealthier ones.

Table 5. Average cost difference and Wilcoxon signed-rank test, by treatment.

	Total expenditure			
	Basket 1	Basket 2	Difference	*p*-value
Benchmark	36.44	36.49	0.05	0.65
NutriScore	30.03	29.29	−0.73	0.46
NutriMark	33.94	32.98	−0.96*	0.03
NutriCouleur	30.86	29.76	−1.11***	0.00
SENS	32.83	33.04	0.21	0.90
NutriRepère	33.47	34.01	0.54	0.86

p-values: *<0.05; ***<0.001.

Table 6. Difference-in-difference estimation of the expenditure associate with FSA improvement.

	Δexpenditure		
	Estimate	Standard error	p-value
(Intercept)	0.03	0.48	0.95
ΔFSA	−0.19	0.25	0.45
NutriScore	−0.46	0.83	0.58
NutriCouleur	−0.85	0.77	0.27
NutriRepère	0.69	0.73	0.35
NutriMark	−0.40	0.77	0.60
SENS	0.34	0.73	0.64
ΔFSA × NutriScore	0.07	0.30	0.81
ΔFSA × NutriCouleur	−0.02	0.34	0.95
ΔFSA × NutriRepère	−0.03	0.35	0.92
ΔFSA × NutriMark	−0.13	0.31	0.68
ΔFSA × SENS	0.02	0.33	0.94

For clarity, the FSA score sign has been inverted.
Significance thresholds: p-values †<0.10; *<0.05 ; **<0.01; ***<0.001.

6.3. Behavioural results

The laboratory context and the computerised shopping platform allow us to shed some light on the behavioural changes that underlie the nutritional and economic results.

6.3.1. Number of products

Over all treatments, neither basket changed significantly in size across label treatments, with the exception of the benchmark in which subjects bought significantly more items than in all other treatments (Table 7, first two rows). This confirms that subjects complied with the instructions. We can exclude boredom as a source of bias for the second basket. The number of items across baskets was not significantly different for any treatment but NC and NM (at 10 per cent) – an indicator that subjects did not respond to labels by simply buying less. Despite the absolute number of items in each basket being mostly constant, subjects did perform some substitutions. On an average, a quarter of the products in each basket was substituted between baskets 1 and 2, with each treatment seeing 4–5 products exiting or entering the basket out of a total of around 20 products (Table 7, second row). Again, this is an indication that subjects did take into account the information displayed by the labels. The fact that even in the baseline treatment, with no labels, subjects substituted roughly a quarter of all products is a sign that subjects varied their choices between the two baskets, by choosing in the absence of any change a different menu out of their habits for the second basket.[7]

6.3.2. Use of back-of-pack information

We recorded each time subjects visualised the ingredients list or the nutritional table (Table 7, third and fourth rows). Usually available on the back of packages in grocery shops, subjects could access such information by clicking on a button on the product screens. For basket 1, subjects looked on average between one and two times at the ingredient lists, and between 0.5 and 1 at nutritional tables, depending on the treatment. Ingredient view frequency significantly decreased in basket 2 for all treatments, including the benchmark; nutrition views decreased for all treatments but the benchmark. On the one hand, this is a straightforward knowledge effect: having acquired the needed information with the first basket, subjects no longer need it for the second. The best estimation of this effect is given by the benchmark treatment, in which no further information was given between baskets. The average number of ingredient and nutrition views decreased by 50 per cent and 32 per cent (not significant), respectively.

On the other hand, the different rates of reduction across labels signal how much labels crowd out the information contained in the nutritional table and ingredients list. Labels can crowd out back-of-pack information both because they contain it already (as in the case of NM, NC and NR, which basically reproduce the nutritional table) or because they summarise it in a focal and credible way (as is the case of NS and SENS). Thus, the drop in the rate of ingredient and nutrition views is an indicator of both the amount of information given by the label and the trust subjects have in that label. All labels crowd out additional information, but to different extents. NR and NC are nearly perfect substitutes of back-of-pack information and as a result lead to a massive decrease in views (between 82 and 92 per cent). At the other end, SENS does result in a smaller (−37 per cent) decrease that is not significantly different from the benchmark (Wilcoxon rank-sum test, p-value = 0.29). This might mean that either SENS fails to give nutritional information or that subjects do not trust the label enough to forfeit looking at the nutritional table.

6.3.3. Label-specific behavioural change

Labels have different granularity, focality and aims. It is, hence, possible to analyse the behavioural change of subjects with respect to the characteristics of each label. Figure 3 summarises our indicators of label-specific behavioural change. While these indicators are not strictly comparable, focusing on them allows us to assess how subjects used the cues given by the labels, and to extract some regularities across treatments.

7 Note that the number of product changes is not directly comparable to the difference in the average number of products. This is because the numbers reported are averages of individual values.

Table 7. Number of clicks, products, nutrition and ingredient views by treatment. Individual averages.

		Benchmark	NS	NC	NR	NM	SENS
No. of products	Basket 1	21.20***	17.90	18.20	19.90	19.80	19.40
	Basket 2	21.00***	17.50	17.60	19.90	19.20	19.40
	Abs difference	−0.30	−0.30	−0.60***	0.00	−0.60†	0.00
No. of product changes	Mean entry	5.30	4.58	3.67	5.80	4.64	4.27
	Mean exit	5.63	4.90	4.20	5.20	4.99	4.22
No. of ingredient views	Basket 1	1.06	0.88	1.89	1.18	0.90	1.45
	Basket 2	0.53	0.16	0.35	0.10	0.16	0.45
	Percentage difference	−49.6%***	−82.3%***	−81.6%***	−91.5%***	−82.2%***	−68.9%***
No. of nutrition views	Basket 1	0.64	0.57	1.14	0.77	0.48	0.73
	Basket 2	0.43	0.17	0.10	0.06	0.13	0.46
	Percentage difference	−31.73%	−69.96%***	−90.83%***	−91.74%***	−72.17%***	−37.04%**

Tests: across treatments Wilcoxon rank-sum test; between baskets by treatment: Wilcoxon signed-rank test; significance thresholds: *p*-values †<0.10; *<0.05; **<0.01; ***<0.001.

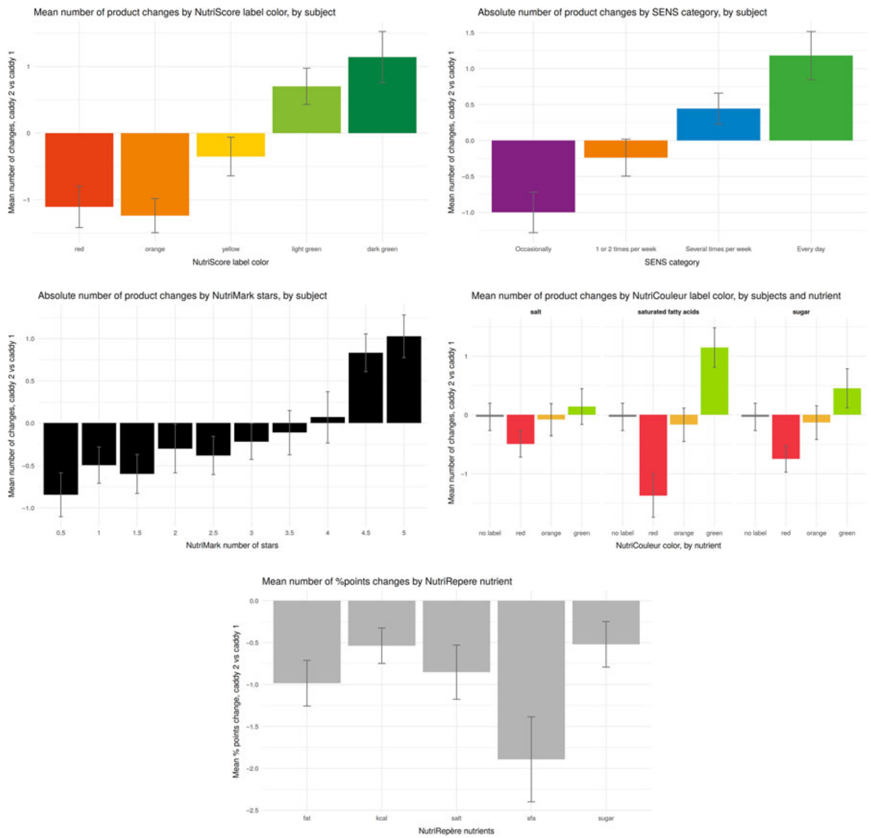


Figure 3. Label-specific behavioural change.

For aggregated labels (NM, SENS and NS), subjects are highly compliant. Nonetheless, they have a tendency to oversimplify the already simple messages carried by the labels by limiting their attention to extreme values or neglecting the nuances of labels. For NM, only the products in the first two categories (five and four and a half stars) are mostly chosen in basket 2 to replace products in basket 1. For SENS, while green products (for every day) replace purple products (occasionally), the two intermediate categories generate little change. For NS, the average number of replaced red and orange products is virtually the same, and light and dark green are very close: despite the fact that the scale has five values, subjects behave as if it had three. Across all aggregated labels, subjects transform more nuanced messages into 'good/bad (or neutral)' signals.

Subjects are also compliant with analytical labels (NC and NR) but to varying degrees depending on nutrients. Subjects pay most attention to fat, some to sugar and not much to salt. Overall, products with red nutrients decrease and products with green nutrients increase with NC. Similarly, the frequency of products with lower bad nutrient content decreases with NR. Given the continuous nature of NR, in Figure 3, we show the average change in percentage reference intake points across the two baskets. Subjects seem to favour fat and saturated fatty acid information.

6.3.4. Label comprehension and assessment

At the end of the experiment, all subjects, except those who took part in the benchmark treatment, completed a questionnaire on the qualitative assessment of the label. Subjects only assessed the label with which they were confronted during the experiment. Accordingly, results do not reflect a *comparative*, but rather *absolute* assessment. Each response was on a four-item Likert scale ('yes', 'mostly', 'not really' and 'no'). Results are given in Table 8. Table 8 reports the summary indicator that is computed by giving scores of 1, 0.75, 0.25 and 0 to each of the four possible answers. Results are robust to changing the values of the intermediate options. Bold indicates the label that has the higher aggregate score for each question.

While differences are small, some patterns are recognisable. NutriScore ranks first for simplicity, usefulness and is considered influential and useful in building better diets that follow nutritional recommendations. On the other end of the spectrum, SENS is the label that most appears as being just an 'advertisement stunt' for products and is ranked last according to several criteria.

7. Discussion

Results indicate that FOPLs can be an effective policy instrument to guide consumers towards healthier diets. Nutritional effects are significant and surprisingly large, especially for the aggregate, colour-coded labels like NutriScore. If nutritional changes of the magnitude recorded in this

Table 8. Qualitative label assessment, aggregate score by treatment.

	NS	NC	NR	NM	SENS
The nutritional label is…					
easy to understand	**0.94**	0.89	0.86	0.91	0.91
useful	**0.94**	0.91	0.88	0.89	0.85
precise	0.71	0.76	0.74	**0.79**	0.57
reassuring	**0.75**	0.70	0.69	0.74	0.67
a tip for good choices	0.73	0.65	0.61	**0.74**	0.65
just advertisement	0.30	0.36	0.34	0.34	**0.41**
The nutritional label does…					
give information about food items to limit	0.79	0.72	0.74	**0.81**	0.77
give information on the nutritional composition	0.60	**0.75**	0.68	0.71	0.46
show me the nutritional quality	**0.79**	**0.79**	**0.79**	**0.79**	0.69
The nutritional label will…					
influence my shopping	**0.71**	0.65	0.67	**0.71**	0.63
help build a better diet	**0.81**	0.72	0.73	0.76	0.62
help in following health recommendations	**0.81**	0.76	0.78	0.78	0.71

experiment were to occur out of the laboratory and survive in the long run, the incidence of nutrition-related diseases would be substantially reduced (see epidemiological evidence as detailed in Donnenfeld *et al.*, 2015; Julia *et al.*, 2015a; Adriouch *et al.*, 2016). There is no significant impact on total expenditure – no label sees a significant expenditure increase for unit nutritional gain. Thus, clear, focal, aggregated labels prove able to generate shopping changes that lead to nutritionally improved choices without inducing an economic cost.

Our results are in line with other findings in the literature. We confirm that simpler, aggregate labels perform better than analytic, detailed labels. We confirm the importance of colour but also show that aggregate, clear information without colour can be effective, since NutriMark, which is black-and-white, performs nearly as well as NutriScore, which is colour coded. Our results are arguably stronger than those usually found in the literature. First, our design allows to observe subjects' compose daily baskets rather than just evaluating a limited set of products. Second and foremost, since we give strong focality to the labels and give plenty of time and information, the detailed, analytical labels should in principle perform better than the simplified designs. We provide the first result in the literature on frequency-based labels (SENS) and show that the results of this alternative approach do not look promising, as the label underperforms the product-based aggregated labels and in some respects also the nutrition-based, analytic ones.

Nonetheless, the external validity of these results needs to be taken with caution. Although we have endeavoured to create a context as close as possible to a real-purchasing situations (sample of regular shoppers; sale at market prices of a wide range of real products; computer interface very close to online or drive-through shopping websites etc.), a crucial difference with real shopping remains: in our setting, labels are highly focal. We intentionally cleared the field of problems of label exposure (i.e. subjects failing to notice the label), understanding (i.e. subject not knowing how to interpret the label) and adoption (i.e. firms might decide not to adopt the label). Baskets 1 and 2 only differ in the presence of labels, and those are explained at length in the instructions given in between the two shopping baskets, and adoption was universal by design. As a result, the subjects' attention is captured by the labels, and understanding problems are minor if not absent. While this feature reinforces the internal validity of our experiment, it is likely to exaggerate the absolute impact of labels compared to real-world situations where consumer attention is limited. It is, therefore, likely that we identify here an upper bound of the potential effect of the labels.

On the other hand, it is possible to argue otherwise. Labels might in the long-run interact with the processes of habit formation (Verplanken and Wood, 2006; Zhen *et al.*, 2011; Daunfeldt, Nordström and Thunström, 2012), and their effect might hence be larger in the long run (which we cannot measure in the laboratory) than in the short run. Moreover, choices are usually made in a social context, and the willingness to signal compliance or superior eating habits could be a further force inflating the effects of the label that we do not capture in our laboratory, in which full privacy is granted (Teyssier, Etilé and Combris, 2014; Etilé and Teyssier, 2016).

A second potential problem with laboratory experiments lies in experimenter demand effects. The aim of our experiment – testing nutritional labels – was transparent to subjects by the time they had to compose the second basket. Moreover, subjects knew that our study had been financed by the Ministry for Health, and a political debate about labels was running in France roughly in parallel with our experimental campaign. It is possible that subjects could show appreciation to the experimenters' work by complying with the labels. Moreover, politically motivated subjects could have inflated their reaction to labels to support their political agenda.

Nonetheless, we believe these effects on the one hand to be minor and on the other hand to be controlled by our design. Recent evidence shows that experimental demand effects are likely low (De Quidt, Haushofer and Roth, 2018; Mummolo and Peterson, 2019) and crowded out by incentives. In our experiment, subjects faced real-purchasing decisions, and they had to buy (a subset of) the products they chose; the cost of sending a political message was hence relatively high. Moreover, to the extent that demand effects and self-serving bias are not label-specific, their impact is fully controlled for in our difference in difference estimations.

While remaining cautious about the absolute magnitude of the labels' impact, the experiment's magnifying glasses allow us to emphasise their relative effects. Having sidestepped the problems of attention, label understanding and label adoption, our experiment can focus on label performance in a world in which consumers are fully informed, labels are fully focal and adoption is universal. As long as potential biases introduced by our design are not treatment specific, our experiment cleanly identifies the relative effect of the labels. Hence, the key result of our study is that colour-coded, aggregate, easy to understand labels have stronger nutritional impacts than analytic, detailed, numeric labels; and that the nutritional adjustment comes at no extra cost for the consumer. This core result is strong and likely to survive intact outside of the laboratory, especially as the labels that are more likely to be taken into consideration and understood are the simpler, colour-coded labels that even in the level playing field of the laboratory win the race.

The best policy tool from a public health perspective appears to be a label that gives synthetic, relevant and focal information, and that can win the trust of consumers. It should be able to stand out among competing stimuli, to be simple to understand and easy to include in each consumers' choice function; it need not necessarily give accurate nutritional information, as long as it is trusted by the consumers to do so. If simple and direct labels work best in a laboratory setting in which subjects had the time and incentive to ponder their choices, we expect them to *a fortiori* work in the real world, where attention, time and cognitive resources are limited.

Disclaimer

Part of the data presented here have already been the object of a short publication with strict nutritional focus, in French, as Crosetto *et al.* (2017). This paper, besides being in English, contains an updated and enlarged nutritional analysis, while the economic and behavioural analyses are entirely new.

Supplementary data

Supplementary data are available at *European Review of Agricultural Economics* online.

Acknowledgements

The study was financed by the French Ministry of Health and Social Affairs as part of the scientific tests ran prior to the formal choice of an official French labelling scheme. The authors would like to thank Marie Cronfalt-Godet, Mariane Damois, Anne Lacroix, Jean-Loup Dupuis and Sabine Pique for invaluable technical and logistic support before and during the experiments, as well as participants to seminars in Parma, Paris, Göttingen, for valuable comments. All errors are the authors'.

References

Adriouch, S., Julia, C., Kesse-Guyot, E., Méjean, C., Ducrot, P., Péneau, S., Donnenfeld, M., Deschasaux, M., Menai, M., Hercberg, S., et al. (2016). Prospective association between a dietary quality index based on a nutrient profiling system and cardiovascular disease risk. *European Journal of Preventive Cardiology* 23: 1669–1676.

Aprile, M. C., Caputo, V. and Nayga, J. R. M. (2012). Consumers' valuation of food quality labels: the case of the European geographic indication and organic farming labels. *International Journal of Consumer Studies* 36: 158–165. doi:10.1111/j.1470-6431.2011.01092.x.

Aschemann-Witzel, J., Grunert, K. G., Trijp, H. C., van Bialkova, S., Raats, M. M., Hodgkins, C., Wasowicz-Kirylo, G. and Koenigstorfer, J. (2013). Effects of nutrition label format and product assortment on the healthfulness of food choice. *Appetite* 71: 63–74.

Australia, N. H. F. O. (1989). The hearth foundation tick. Technical Report, National Heart Foundation of Australia.

Australia, C. O. (2013a). Proposed front-of-pack food labeling designs: qualitative research outcomes. Technical Report, Hall & Partners Open Mind.

Australia, C. O. (2013b). Proposed Front-of-pack food labeling designs: quantitative research outcomes. Technical Report, Hall & Partners Open Mind.

Balcombe, K., Fraser, I. and Falco, S. D. (2010). Traffic lights and food choice: a choice experiment examining the relationship between nutritional food labels and price. *Food Policy* 35: 211–220. doi:10.1016/j.foodpol.2009.12.005.

Becker, M. W., Bello, N. M., Sundar, R. P., Peltier, C. and Bix, L. (2015). Front of pack labels enhance attention to nutrition information in novel and commercial brands. *Food Policy* 56: 76–86. doi:10.1016/j.foodpol.2015.08.001.

Berning, J. P., Chouinard, H. H. and McCluskey, J. J. (2011). Do positive nutrition shelf labels affect consumer behavior? Findings from a field experiment with scanner data. *American Journal of Agricultural Economics* 93: 364–369. doi:10.1093/ajae/aaq104.

Bialkova, S. and Van Trijp, H. (2010). What determines consumer attention to nutrition labels? *Food Quality and Preference* 21: 1042–1051. doi:10.1016/j.foodqual.2010.07.001.

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: 184.

Boztuğ, Y., Juhl, H. J., Elshiewy, O. and Jensen, M. B. (2015). Consumer response to monochrome guideline daily amount nutrition labels. *Food Policy* 53: 1–8. doi:10.1016/j.foodpol.2015.03.002.

Cecchini, M. and Warin, L. (2016). Impact of food labelling systems on food choices and eating behaviours: a systematic review and meta-analysis of randomized studies. *Obesity Reviews* 17: 201–210.

Cooper, S., Pelly, F. and Lowe, J. (2017). Assessment of the construct validity of the Australian health star rating: a nutrient profiling diagnostic accuracy study. *European Journal of Clinical Nutrition* 71: 1353–1359.

Council of the European Union (2011). Regulation (EU) no. 1169/2011. http://data.europa.eu/eli/reg/2011/1169/oj.

Crockett, R. A., Jebb, S. A., Hankins, M. and Marteau, T. M. (2014). The impact of nutritional labels and socioeconomic status on energy intake. An experimental field study. *Appetite* 81: 12–19. doi:10.1016/j.appet.2014.05.024.

Crosetto, P., Lacroix, A., Muller, L. and Ruffieux, B. (2017). Modification des achats alimentaires en réponse à cinq logos nutritionnels. *Cahiers de Nutrition et de Diététique* 5262: 129–133. doi:10.1016/j.cnd.2017.04.002.

Crosetto, P., Muller, L. and Ruffieux, B. (2016a). 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. *Journal of Economic Psychology* 55: 30–50. doi:10.1016/j.joep.2016.03.006.

Crosetto, P., Muller, L. and Ruffieux, B. (2016b). Réponses des consommateurs à trois systèmes d'étiquetage nutritionnel face avant. *Cahiers de Nutrition et de Diététique* 5011: 111–160. doi:10.1016/j.cnd.2016.04.002.

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

Daunfeldt, S.-O., Nordström, J. and Thunström, L. (2012). Habit formation in food consumption. In: J. L. Lusk, J. Roosen and J. F. Shogren (eds), *The Oxford Handbook of the Economics of Food Consumption and Policy*. Oxford: Oxford Handbooks Online.

De Quidt, J., Haushofer, J. and Roth, C. (2018). Measuring and bounding experimenter demand. *American Economic Review* 108: 3266–3302. doi:10.1257/aer.20171330.

Donnenfeld, M., Julia, C., Kesse-Guyot, E., Méjean, C., Ducrot, P., Péneau, S., Deschasaux, M., Latino-Martel, P., Fezeu, L., Hercberg, S., et al. (2015). Prospective association between cancer risk and an individual dietary index based on the British Food Standards Agency Nutrient Profiling System. *British Journal of Nutrition* 114: 1702–1710.

Drewnowski, A. and Darmon, N. (2005). The economics of obesity: dietary energy density and energy cost. *The American Journal of Clinical Nutrition* 82: 265S–273S. doi:10.1093/ajcn/82.1.265S.

Drewnowski, A. and Eichelsdoerfer, P. (2010). Can low-income Americans afford a healthy diet? *Nutrition Today* 44: 246.

Drewnowski, A. and Specter, S. (2004). Poverty and obesity: the role of energy density and energy costs. *The American Journal of Clinical Nutrition* 79: 6–16. doi:10.1093/ajcn/79.1.6.

Drichoutis, A. C., Lazaridis, P., Nayga, R. M., Kapsokefalou, M. and Chryssochoidis, G. (2008). A theoretical and empirical investigation of nutritional label use. *The European Journal of Health Economics* 9: 293–304.

Ducrot, P., Julia, C., Méjean, C., Kesse-Guyot, E., Touvier, M., Fezeu, L. K., Hercberg, S. and Péneau, S. (2016). Impact of different front-of-pack nutrition labels on consumer purchasing intentions: a randomized controlled trial. *American Journal of Preventive Medicine* 50: 627–636.

Ducrot, P., Méjean, C., Julia, C., Kesse-Guyot, E., Touvier, M., Fezeu, L. K., Hercberg, S. and Péneau, S. (2015). Objective understanding of front-of-package nutrition labels among nutritionally at-risk individuals. *Nutrients* 7: 7106–7125.

Egnell, M., Talati, Z., Hercberg, S., Pettigrew, S. and Julia, C. (2018). Objective understanding of front-of-package nutrition labels: an international comparative experimental study across 12 countries. *Nutrients* 10: 1542, doi:10.3390/nu10101542.

Etilé, F. and Teyssier, S. (2016). Signaling corporate social responsibility: third-party certification versus brands. *The Scandinavian Journal of Economics* 118: 397–432.

Feunekes, G. I., Gortemaker, I. A., Willems, A. A., Lion, R. and Kommer, M. van den (2008). Front-of-pack nutrition labelling: testing effectiveness of different nutrition labelling formats front-of-pack in four European countries. *Appetite* 50: 57–70.

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.

Gaigi, H., Raffin, S., Maillot, M., Adrover, L., Ruffieux, B. and Darmon, N. (2015). Expérimentation d'un fléchage nutritionnel dans deux supermarchés à Marseille «Le choix Vita+». *Cahiers de Nutrition et de Diététique* 50: 16–24. doi:10.1016/j.cnd.2014.12.005.

Gracia, A., Loureiro, M. L. and Nayga, R. M. (2009). Consumers' valuation of nutritional information: a choice experiment study. *Food Quality and Preference* 20: 463–471. doi:10.1016/j.foodqual.2009.03.010.

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: 385–399. doi:10.1007/s10389-007-0101-9.

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

Haut Conseil de la Santé Publique (2015). Avis relatif à l'information sur la qualité nutritionnelle des produits alimentaires. Avis du 25 juin 2015, Haut Conseil de la Santé Publique.

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 Health Nutrition* 16: 430–439. doi:10.1017/S1368980012000754.

Helfer, P. and Shultz, T. R. (2014). The effects of nutrition labeling on consumer food choice: a psychological experiment and computational model. *Annals of the New York Academy of Sciences* 1331: 174–185. doi:10.1111/nyas.12461.

Hersey, J. C., Wohlgenant Kelly, C., Arsenault Joanne, E., Kosa Katherine, M. and Muth Mary, K. (2013). Effects of front-of-package and shelf nutrition labeling systems on consumers. *Nutrition Reviews* 71: 1–14. doi:10.1111/nure.12000.

Huang, L. and Lu, J. (2016). The impact of package color and the nutrition content labels on the perception of food healthiness and purchase intention. *Journal of Food Products Marketing* 22: 191–218. doi:10.1080/10454446.2014.1000434.

Imram, N. (1999). The role of visual cues in consumer perception and acceptance of a food product. *Nutrition & Food Science* 99: 224–230. doi:10.1108/00346659910277650.

Iyengar, S. S. and Lepper, M. R. (2000). When choice is demotivating: can one desire too much of a good thing? *Journal of Personality and Social Psychology* 79: 995–1006.

Julia, C., Blanchet, O., Méjean, C., Péneau, S., Ducrot, P., Allès, B., Fezeu, L. K., Touvier, M., Kesse-Guyot, E., Singler, E. and Hercberg, S. (2016). Impact of the front-of-pack 5-colour nutrition label (5-CNL) on the nutritional quality of purchases: an experimental study. *International Journal of Behavioral Nutrition and Physical Activity* 13: 101, doi:10.1186/s12966-016-0416-4.

Julia, C., Ducrot, P., Lassale, C., Fézeu, L., Méjean, C., Péneau, S., Touvier, M., Hercberg, S. and Kesse-Guyot, E. (2015a). Prospective associations between a dietary index based on the British Food Standard Agency nutrient profiling system and 13-year weight gain in the SU.VI.MAX cohort. *Preventive Medicine* 81: 189–194.

Julia, C. and Hercberg, S. (2017). Development of a new front-of-pack nutrition label in France: the five-colour nutri-score. *Public Health Panorama* 3: 712–725.

Julia, C., Kesse-Guyot, E., Ducrot, P., Péneau, S., Touvier, M., Méjean, C. and Hercberg, S. (2015b). Performance of a five category front-of-pack labelling system – the 5-colour nutrition label – to differentiate nutritional quality of breakfast cereals in France. *BMC Public Health* 15: 179, doi:10.1186/s12889-015-1522-y.

Julia, C., Kesse-Guyot, E., Touvier, M., Méjean, C., Fezeu, L. and Hercberg, S. (2014). Application of the British Food Standards Agency nutrient profiling system in a French food composition database. *British Journal of Nutrition* 112: 1699–1705. doi:10.1017/S0007114514002761.

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: 120–129. doi:10.1093/heapro/dap012.

Kivetz, R. and Simonson, I. (2000). The effects of incomplete information on consumer choice. *Journal of Marketing Research* 37: 427–448. doi:10.1509/jmkr.37.4.427.18796.

Kleef, E. V. and Dagevos, H. (2015). The growing role of front-of-pack nutrition profile labeling: a consumer perspective on key issues and controversies. *Critical Reviews in Food Science and Nutrition* 55: 291–303.

Loo, E. J. V., Caputo, V., Nayga, R. M., Meullenet, J.-F. and Ricke, S. C. (2011). Consumers' willingness to pay for organic chicken breast: evidence from choice experiment. *Food Quality and Preference* 22: 603–613. doi:10.1016/j.foodqual.2011.02.003.

Lusk, J. L. and Shogren, J. F. (2007). *Experimental Auctions: Methods and Applications in Economic and Marketing Research*. Cambridge: Cambridge University Press.

Malhotra, N. K. (1982). Information load and consumer decision making. *Journal of Consumer Research* 8: 419–430. 00581.

Marchi, E. D., Caputo, V., Nayga, R. M. and Banterle, A. (2016). Time preferences and food choices: evidence from a choice experiment. *Food Policy* 62: 99–109. doi:10.1016/j.foodpol.2016.05.004.

Ministère des Affaires Sociales et de la Santé, Fonds Français pour l' Alimentation et la Santé (2017). Evaluation ex ante, en conditions reéelles d'achat, de divers systèmes graphiques d'eétiquetage nutritionnel – rapport final du comité scientifique.

Muller, L., Lacroix, A., Lusk, J. L. and Ruffieux, B. (2017). Distributional impacts of fat taxes and thin subsidies. *The Economic Journal* 127: 2066–2092. doi:10.1111/ecoj.12357.

Muller, L. and Prevost, M. (2016). What cognitive sciences have to say about the impacts of nutritional labelling formats. *Journal of Economic Psychology* 55: 17–29. doi:10.1016/j.joep.2016.01.005.

Muller, L. and Ruffieux, B. (2012). Modification des achats en réponse à l'apposition de différents logos d'évaluation nutritionnelle sur la face avant des emballages. *Cahiers de Nutrition et de Diététique* 47: 171–182. doi:10.1016/j.cnd.2012.02.004.

Mummolo, J. and Peterson, E. (2019). Demand effects in survey experiments: an empirical assessment. *American Political Science Review* 113: 517–529. doi:10.1017/S0003055418000837.

Mørk, T., Grunert, K. G., Fenger, M., Juhl, H. J. and Tsalis, G. (2017). An analysis of the effects of a campaign supporting use of a health symbol on food sales and shopping behaviour of consumers. *BMC Public Health* 17: 239.

Neal, B., Crino, M., Dunford, E., Gao, A., Greenland, R., Li, N., Ngai, J., Ni Mhurchu, C., Pettigrew, S., Sacks, G., et al. (2017). Effects of different types of front-of-pack labelling information on the healthiness of food purchases – a randomised controlled trial. *Nutrients* 9: 1284.

Norden (2010). The Keyhole: healthy choices made easy. Technical Report, Nordic Cooperation – Nordic Council of Ministries.

Nordström, J. and Thunström, L. (2011a). Can targeted food taxes and subsidies improve the diet? Distributional effects among income groups. *Food Policy* 36: 259–271.

Nordström, J. and Thunström, L. (2011b). Economic policies for healthier food intake: the impact on different household categories. *The European Journal of Health Economics* 12: 127–140. doi:10.1007/s10198-007-0052-7.

Noussair, C., Robin, S. and Ruffieux, B. (2004). Do consumers really refuse to buy genetically modified food? *The Economic Journal* 114: 102–120.

Nugier, A., Serry, A. and Thanh, V. (2016). Système d'information nutritionnelle à cinq couleurs sur la face avant des emballages alimentaires: comparaison de trois formats de logo. *Bulletin Epidémiologique Hebdomadaire* 34: 598–607.

Prevost, M., Hot, P., Muller, L., Ruffieux, B., Cousin, E., Pichat, C. and Baciu, M. (2017). Neural correlates of the healthiness evaluation processes of food labels. *Nutritional Neuroscience* 0: 1–11. doi:10.1080/1028415X.2017.1309820.

Rao, M., Afshin, A., Singh, G. and Mozaffarian, D. (2013). Do healthier foods and diet patterns cost more than less healthy options? A systematic review and meta-analysis. *BMJ Open* 3: e004277, doi:10.1136/bmjopen-2013-004277.

Rayner, M., Scarborough, P. and Lobstein, T. (2009). *The UK Ofcom Nutrient Profiling Model: Defining 'Healthy'and 'Unhealthy' Foods and Drinks for TV Advertising to Children*. London: OfCom.

Rayner, M., Scarborough, P. and Williams, C. (2004). The origin of guideline daily amounts and the food standards agency's guidance on what counts as 'a lot' and 'a little'. *Public Health Nutrition* 7: 549–556.

Roberto, C. A., Bragg, M. A., Schwartz, M. B., Seamans, M. J., Musicus, A., Novak, N. and Brownell, K. D. (2012a). Facts up front versus traffic light food labels. *American Journal of Preventive Medicine* 43: 134–141. doi:10.1016/j.amepre.2012.04.022.

Roberto, C. A., Shivaram, M., Martinez, O., Boles, C., Harris, J. L. and Brownell, K. D. (2012b). The Smart Choices front-of-package nutrition label. Influence on perceptions and intake of cereal. *Appetite* 58: 651–657. 10.1016/j.appet.2012.01.003.

République, F. (2017). Arrêté du 31 octobre 2017 fixant la forme de présentation complémentaire à la déclaration nutritionnelle recommandée par l'etat en application des articles l. 3232–8 et r. 3232–7 du code de la santé publique. *Journal Officiel de la République Française* 0257.

Scheibehenne, B., Miesler, L. and Todd, P. M. (2007). Fast and frugal food choices: uncovering individual decision heuristics. *Appetite* 49: 578–589.

Schuldt, J. P. (2013). Does green mean healthy? Nutrition label color affects perceptions of healthfulness. *Health Communication* 28: 814–821. doi:10.1080/10410236.2012.725270.

Schulte-Mecklenbeck, M., Sohn, M., Bellis, E., de Martin, N. and Hertwig, R. (2013). A lack of appetite for information and computation. Simple heuristics in food choice. *Appetite* 71: 242–251. doi:10.1016/j.appet.2013.08.008.

Schwarz, B. (2004). *The Paradox of Choice: Why More Is Less: How the Culture of Abundance Robs Us of Satisfaction*. New York, NY: Harper Collins.

Shogren, J. F. (2011). *The Oxford Handbook of the Economics of Food Consumption and Policy*. Oxford: Oxford University Press.

Spanos, S., Kenda, A. S. and Vartanian, L. R. (2015). Can serving-size labels reduce the portion-size effect? A pilot study. *Eating Behaviors* 16: 40–42. doi:10.1016/j.eatbeh.2014.10.007.

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

Teyssier, S., Etilé, F. and Combris, P. (2014). Social-and self-image concerns in fair-trade consumption. *European Review of Agricultural Economics* 42: 579–606.

Tharrey, M., Maillot, M., Azaïs-Braesco, V. and Darmon, N. (2017). From the SAIN, LIM system to the SENS algorithm: a review of a French approach of nutrient profiling. *The Proceedings of the Nutrition Society* 76: 237–246. doi:10.1017/S0029665117000817.

Thunström, L. and Nordström, J. (2015). Determinants of food demand and the experienced taste effect of healthy labels – an experiment on potato chips and bread. *Journal of Socio-Economics* 56: 13–20. doi:10.1016/j.socec.2015.02.004.

Variyam, J. N. (2008). Do nutrition labels improve dietary outcomes? *Health Economics* 17: 695–708. doi:10.1002/hec.1287.

Vasiljevic, M., Pechey, R. and Marteau, T. M. (2015). Making food labels social: the impact of colour of nutritional labels and injunctive norms on perceptions and choice of snack foods. *Appetite* 91: 56–63. doi:10.1016/j.appet.2015.03.034.

Verplanken, B. and Wood, W. (2006). Interventions to break and create consumer habits. *Journal of Public Policy & Marketing* 25: 90–103. 10.1509/jppm.25.1.90.

Viswanathan, M., Torelli, C. J., Xia, L. and Gau, R. (2009). Understanding the influence of literacy on consumer memory: the role of pictorial elements. *Journal of Consumer Psychology* 19: 389–402. doi:10.1016/j.jcps.2009.04.002.

Vlaeminck, P., Jiang, T. and Vranken, L. (2014). Food labeling and eco-friendly consumption: experimental evidence from a Belgian supermarket. *Ecological Economics* 108: 180–190.

Volkova, E. and Mhurchu, C. N. (2015). The influence of nutrition labeling and point-of-purchase information on food behaviours. *Current Obesity Reports* 4: 19–29.

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: 709–720.

Vyth, E. L., Steenhuis, I. H., Vlot, J. A., Wulp, A., Hogenes, M. G., Looije, D. H., Brug, J. and Seidell, J. C. (2010). Actual use of a front-of-pack nutrition logo in the supermarket: consumers' motives in food choice. *Public Health Nutrition* 13: 1882–1889. doi:10.1017/S1368980010000637.

Waterlander, W. E., Steenhuis, I. H., Boer, M. R., de, Schuit, A. J. and Seidell, J. C. (2013). Effects of different discount levels on healthy products coupled with a healthy choice label, special offer label or both: results from a web-based supermarket experiment. *International Journal of Behavioral Nutrition and Physical Activity* 10: 59.

Watson, W. L., Kelly, B., Hector, D., Hughes, C., King, L., Crawford, J., Sergeant, J. and Chapman, K. (2014). Can front-of-pack labelling schemes guide healthier food choices? Australian shoppers' responses to seven labelling formats. *Appetite* 72: 90–97. doi:10.1016/j.appet.2013.09.027.

Withrow, D. and Alter, D. A. (2011). The economic burden of obesity worldwide: a systematic review of the direct costs of obesity. *Obesity Reviews* 12: 131–141. doi:10.1111/j.1467-789X.2009.00712.x.

Wong, C. L., Mendoza, J., Henson, S. J., Qi, Y., Lou, W. and L'Abbé, M. R. (2014). Consumer attitudes and understanding of cholesterol-lowering claims on food: randomize mock-package experiments with plant sterol and oat fibre claims. *European Journal of Clinical Nutrition* 68: 946–952. doi:10.1038/ejcn.2014.107.

Zhen, C., Wohlgenant, M. K., Karns, S. and Kaufman, P. (2011). Habit formation and demand for sugar-sweetened beverages. *American Journal of Agricultural Economics* 93: 175–193. doi:10.1093/ajae/aaq155.