

Manuscript version: Author's Accepted Manuscript

The version presented in WRAP is the author's accepted manuscript and may differ from the published version or Version of Record.

Persistent WRAP URL:

<http://wrap.warwick.ac.uk/156440>

How to cite:

Please refer to published version for the most recent bibliographic citation information.

Copyright and reuse:

The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions.

Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

Publisher's statement:

Please refer to the repository item page, publisher's statement section, for further information.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk.

Computational Methods for Predicting and Understanding Food Judgment

Gandhi, N.* , Zou, W.^, Meyer, C.^, Bhatia, S.* & Walasek, L.^

^ University of Warwick

* University of Pennsylvania

16/07/2021

In Press at Psychological Science

Gandhi and Zou are joint first authors. Address correspondence to Lukasz Walasek at

L.Walasek@warwick.ac.uk.

Abstract

People make subjective judgments about the healthiness of different foods every day, which in turn influence their food choices and health outcomes. Despite their importance, there are few quantitative theories about the psychological underpinnings of such judgments. This study introduces a novel computational approach that can approximate people's knowledge representations for thousands of common foods. We use these representations to predict how both lay decision-makers (general population) and experts judge the healthiness of individual foods. We also apply our method to predict the impact of behavioral interventions such as the provision of front-of-pack nutrient and calorie information. Across multiple studies with data from 846 adults, our models achieve very high accuracy rates (r^2 from 0.65 to 0.77), and significantly outperform competing models based on factual nutritional content. These results illustrate how new computational methods applied to established psychological theory can be used to better predict, understand, and influence health behavior.

Statement of Relevance

Is granola healthy? What about steak? What type of knowledge do we use when judging the healthiness of different foods? To answer this question, we study how different food names tend to co-occur with other words in large-scale language data. We use this information to predict people's judgments of food healthiness and to uncover words and concepts that are more associated with healthy and unhealthy foods. Our results show that people's judgments of food healthiness are largely explained by the strength of association with naturalness and rawness. In a series of experiments, we demonstrate that these associations play a significant role in explaining judgments of healthiness even if people are shown front-of-pack nutrient and calorie information.

Keywords: Judgment; food healthiness perceptions, knowledge representations; word embedding; food labeling; computational models

Computational Methods for Predicting and Understanding Food Judgment

Poor diet is one of the primary preventable causes of premature death in high-income countries (Bauer et al., 2014). Understandably, people want to consume healthy foods as they recognize the relationship between diet and health. However, people can only make healthy food choices to the extent that they can correctly judge a food's healthiness. One obstacle to healthy eating is that there is no normative answer to the question: "what makes food healthy or unhealthy?" (Lobstein & Davies, 2008). Still, it is commonly believed that food healthiness judgments are strongly linked to beliefs about the nutritional content of food products (Scarborough et al., 2007).

Indeed, health organizations worldwide routinely emphasize which nutrients people should avoid (high saturates, fats, sugars, salt) and which they should consume more of (high protein, fiber) (Lobstein & Davies, 2008). This is apparent in the design of numerous front-of-pack food labeling formats, which attempt to simplify complex nutrient information for consumers. Such interventions highlight overall energy content and the presence of nutrients that are most associated with the rising rates of obesity and chronic diseases (Kanter et al., 2018). Yet, evidence about the effectiveness of such interventions is mixed (Sanjari et al., 2017).

The success of front-of-pack labeling rests on the assumption that people rely on energy and nutrient information to judge a food's healthiness (Orquin, 2014). However, evidence suggests that healthiness judgments reflect pre-existing knowledge that people associate with foods' perceived naturalness (Siipi, 2012) and taste (Turnwald et al., 2017). These are further influenced by cultural traditions (Pieniak et al., 2009), previous eating experiences (Papies, 2013), media/advertisements (Whalen et al., 2018), background nutrition knowledge (Soederberg Miller & Cassady, 2015), choice context (Downs et al., 2015), product category (Plasek et al., 2020), packaging (Reutner et al., 2015), and health halo effects of labels such as "organic" (Perkovic & Orquin, 2018; Schuldt & Schwarz, 2010).

These factors contribute to the diverse and multidimensional knowledge representations that decision-makers draw upon when making food-related judgments and choices. Indeed, specific knowledge representations that are retrieved from memory (Scheibehenne et al., 2007) or explicitly provided to the decision-maker (Schulte-Mecklenbeck et al., 2013) can be used to make choices between food items using simple heuristics. Whereas knowledge representations may explain why people think some foods are healthier than others, they may be biased, causing systematic and

PREDICTING FOOD HEALTHINESS JUDGMENT

predictable errors in healthiness perception. This could explain why people's judgments of healthiness deviate from an estimate of healthiness based on nutrient and energy values of the food (Orquin, 2014).

Researchers studying food judgment and choice typically rely on knowledge representations that are restricted to a predefined and limited set of factors and attributes (Step toe et al., 1995). This also means that current approaches are not well suited for making generalizable predictions about judgments in the presence of interventions, such as different food labeling strategies (Kanter et al., 2018). How can we identify and quantify knowledge representations that underpin people's judgment of food healthiness? We propose a novel approach to overcome these challenges, which relies on recent advances in computational linguistics. Unlike previous approaches, in which food representations were either manually specified by the researchers or based on self-reports, we establish food representations using natural language data. More specifically, we use word distribution statistics in large text corpora to uncover quantitative representations for words and phrases that describe food items. The use of this type of data means that uncovered representations reflect information conveyed in language, which individuals may use to form beliefs, and may even guide everyday health judgment. We find support for this prediction by studying how knowledge representations retrieved from natural language can account for judgments of food healthiness across six experiments. Our further analysis reveals that our models perform well because they capture associations related to naturalness or rawness of foods.

The knowledge representations used in our analysis are high-dimensional vectors for words (also known as word embeddings) (Landauer & Dumais, 1997; Lenci, 2018; Mikolov et al., 2013). A useful property of word vectors is that the proximities between vectors measure the associations between individual words. These associations have been shown to correlate with human semantic, factual, probability, and social judgments (Bhatia, 2017; Caliskan et al., 2017; Pereira et al., 2016). Recently, researchers have shown that these word vectors can be used to quantify people's knowledge about various natural entities by using these as inputs into regressions that predict more complex (potentially non-associative) judgments in other domains (Bhatia, 2019; Richie et al., 2019; Zou & Bhatia, 2021).

Our approach is as follows: First, we obtain high-dimensional vector representations for food items from popular word embedding models trained on large-scale textual datasets. We hypothesize

PREDICTING FOOD HEALTHINESS JUDGMENT

that these word vectors may serve as a good approximation of knowledge representations that underpin judgments of food healthiness. To test this proposition, with some training data involving people's ratings of diverse food items, we learn a mapping from our high-dimensional vector space to the (one-dimensional) scale that measures perceptions of healthiness (i.e. people's judgments). We then apply this mapping to food items outside of the training data to predict people's judgments for these "out-of-sample" food items. Note that such a mapping identifies regions of the vector space implicitly associated with food healthiness, and thus can be used to understand the conceptual and associative underpinnings of health judgment. We can also build this kind of mapping separately for different groups of people, to predict judgments of both lay and expert judges, as well as differences in judgments between individuals exposed to different front-of-pack labeling strategies. Across six studies, we demonstrate the generalizability, accuracy, and power of our approach.

Studies 1A, 1B and 1C

Our primary objective was to establish the feasibility of our computational approach in predicting people's judgments of food healthiness. Therefore, we elicited judgments of healthiness for a wide range of food items (presented as food names) from the general population (Study 1A) and from a sample of registered dietitians (Study 1B). In Study 1C, we tested the performance of our models on healthiness judgments of foods' names *and* images.

Methods

Participants

Our approach does not rely on standard null hypothesis testing but rather on maximizing out-of-sample predictions. Using previous work for guidance (Bhatia, 2019), we chose to obtain judgments for a diverse set of 172 foods and aimed to recruit at least 100 participants (with each participant judging each of the 172 food items). The only exception was in Study 1B where we prioritized how many responses we could obtain from nutritional experts in a three-month window. Note that in all studies, the primary unit of analysis was the average healthiness rating, across all participants, for a given food item.

In all studies, only participants over 18 years of age were eligible to take part. Our only exclusion criterion was based on the correlation between each person's food ratings and the grand

PREDICTING FOOD HEALTHINESS JUDGMENT

mean of aggregate ratings for those foods within the sample. Prior to data analysis (in all studies reported in this paper), we removed participants with a correlation lower than 0.4 with the grand mean of all ratings in a given study (based on the inspection of data from Study 1A). Although this exclusion criterion aimed to remove participants with very noisy ratings that would generate outlier responses, an analysis of the full sample shows that none of our results are affected by this exclusion criteria (see Section 1 of the Supplementary Materials).

For Study 1A, 149 participants were recruited from Prolific Academic in return for a fixed payment of £1.30. Using the aforementioned criterion, data from 15 respondents were removed leaving 134 participants as our final sample (aged 18-74 years, $M_{\text{age}} = 29.57$ years, $SD = 8.86$, 43% females, and 84% had no dietary restrictions). For Study 1B, we contacted registered dietitians after a formal introduction by email with a request to take part in our study and forward the invitation to their colleagues. We also advertised the study on personal social media accounts. As an incentive, participants were able to request a report of the main findings. Nineteen registered dietitians took part in the study (none excluded, aged 23-56, $M_{\text{age}} = 35.84$ years, $SD = 10.36$, 89% females and 68% had no dietary restrictions). One hundred participants recruited on Prolific Academic took part in Study 1C in return for a fixed payment of £1.90. We excluded one participant based on the same criteria as in Study 1A. This left 99 participants in our final sample (aged 18-69 years, $M_{\text{age}} = 27.25$ years, $SD = 10.20$, 44% females, and 82% had no dietary restrictions). A detailed breakdown of participants' characteristics for this and other studies reported here is provided in the Supplementary Materials (Section 2). This research was approved by the University of Warwick's Biomedical and Scientific Research Ethics Sub-Committee (approval # REGO-2018-2268).

Design and Procedure

In all studies, participants were asked to simply judge the healthiness of 172 foods on a scale ranging from -100 (extremely unhealthy) to +100 (extremely healthy). In Study 1A and 1B, each food was described using its name only. In Study 1C, a generic image of the food item was presented directly below the food name. Responses were made using a slider, with its starting position always at zero by default (neither healthy nor unhealthy). This scale was chosen because it is fine-grained (200 intervals) and balanced (symmetric around 0), offering nearly continuous data for predictive modeling (Bhatia, 2019). Participants had the option of selecting "Don't know" if they were unfamiliar with a food item, with those ratings removed from the analysis. The order of the items was randomized for every

PREDICTING FOOD HEALTHINESS JUDGMENT

participant and only one item was visible at a time. The same generic task instruction: “Using the slider, please use your first impression to rate the following food item according to the scale below:” was displayed above all stimuli in every study condition. After rating all foods, participants were asked about their age, gender, and dietary restrictions (with the options of “Pescetarian (no meat, but eat fish and/or shellfish)”, “Vegetarian”, “Vegan”, “Other (please specify if you wish)” and “None of the above”). Our nutritional experts in Study 1B were also asked two additional demographic questions at the end of the survey (namely, “No. of Years as a Registered Dietitian” and “Area of Specialism”).

Materials

We obtained a list of foods from the USDA Food Composition Database, the most recent official publication of nutrient information pertaining to over 3102 unique food items (U.S. Department of Agriculture, 2019). Only foods present in the vocabulary of the pre-trained word2vec model were considered, leaving 571 food items (see the Computational Approach section for detail). Two hundred food items, across all food categories (e.g. vegetables, meats, dishes), were manually chosen by co-author WZ to ensure diversity in the stimuli set. Next, co-authors NG and LW removed uncommon and ambiguous food items such as “squash” (because of its additional meaning related to sports), resulting in the final list of 172 food items (see the OSF repository associated with this project for the full list: <https://osf.io/jys6u/>). Note that the same list of 172 food items was used in all studies reported in the main text of the manuscript.

In Study 1C, 69 of the food images were directly sourced from an image database for experimental research (Blechert et al., 2019), with the remaining 103 images sourced online and standardized to match (white background, 600 x 450 dimensions, and jpeg format).

Computational approach

We used three statistical models to predict subjective food healthiness judgments. Our analysis relied on participants’ judgments at the aggregate level. We evaluated the accuracy of each of our three statistical models in predicting subjective food healthiness judgments using leave-one-out cross-validation, which means that we trained our models on all but one aggregate judgment (“training data”) and used the trained model to predict the rating of the left-out food item (“test data”). We repeated this procedure 172 times so that each food item was in the test data once. Cross-validation ensures that our modeling avoided overfitting and that performance of each model was evaluated based on model generalizability.

PREDICTING FOOD HEALTHINESS JUDGMENT

In the first model, the *Nutrient Model*, we used nutrient content information to predict healthiness judgments. This model was an ordinary-least-squared regression with main effects for food calorie content, amounts of nutrients (fat, saturates, sugar, salt, and protein) per 100g, and the relative coding scheme based on the UK traffic light labeling for fat, saturates, sugar and salt (green, amber and red). Under the traffic light labeling system, green signifies a healthier food choice to consumers implying “go ahead”; amber indicates the item contains moderate amounts of the negative nutrient(s); and red signals caution for overconsumption (Trudel et al., 2015). The model was fit on the training data, and the best fitting parameters of the model were applied to the nutrient information of the (out-of-sample) food, in order to predict participant ratings. The nutrients and calorie information included in the Nutrient Model reflects the current European Union’s regulations concerning mandatory information for food package labeling (Article 30, Regulation No. 1169/2011 European Commission, 2011). In the Supplementary Materials (Section 3), we summarize tests of the robustness of our results using three extended versions of the Nutrient Model. First, we expanded the Nutrient Model to incorporate the potential role of 23 nutrients (e.g., fiber, calcium, and Vitamin C). Second, we also tested a version of the model that used nutrient amounts per portion size, defined as the amount per 100 calories. We also combined these two extensions into our final, third model.

In our *Vector Representation Model*, we used vector representations from the word2vec model (Mikolov et al., 2013). This model was pre-trained on a dataset of Google News articles, and has 300-dimensional vector representations for three million common words and phrases in the English language (see Mikolov et al., 2013 for details). In designing our studies, we only considered foods whose name features in the pre-trained word2vec model. We also analyzed the predictions of other established pre-trained word vector models, which we discuss in the Supplementary Materials (Section 4). In our main analysis, we used normalized word vectors, in which the magnitude of the vectors was scaled to be equal to 1. We regressed participants’ healthiness ratings on these vectors, which allowed the model to learn a linear mapping from the semantic vector space to health judgments. This learnt mapping was then applied to the vectors of other (out-of-sample) foods, in order to predict participant ratings of those foods, and measure the models’ predictive accuracy.

Because of the high number of predictor variables in this model (300), we applied a regularized regression technique known as ridge regression. Ridge regression allows high numbers of predictors to be considered and takes into account whether predictors are highly correlated. In

PREDICTING FOOD HEALTHINESS JUDGMENT

previous and similar work, ridge regression was the best-fitting regression technique for mapping pre-trained 300-dimensional vector representations to judgments and was consequently chosen for our analysis (Bhatia, 2019; Richie et al., 2019). We also tested other regression techniques including lasso, support vector, and k-nearest neighbors regression and found ridge regression was indeed the best-fitting regression. We discuss this robustness test in the Supplementary Materials (Section 5).

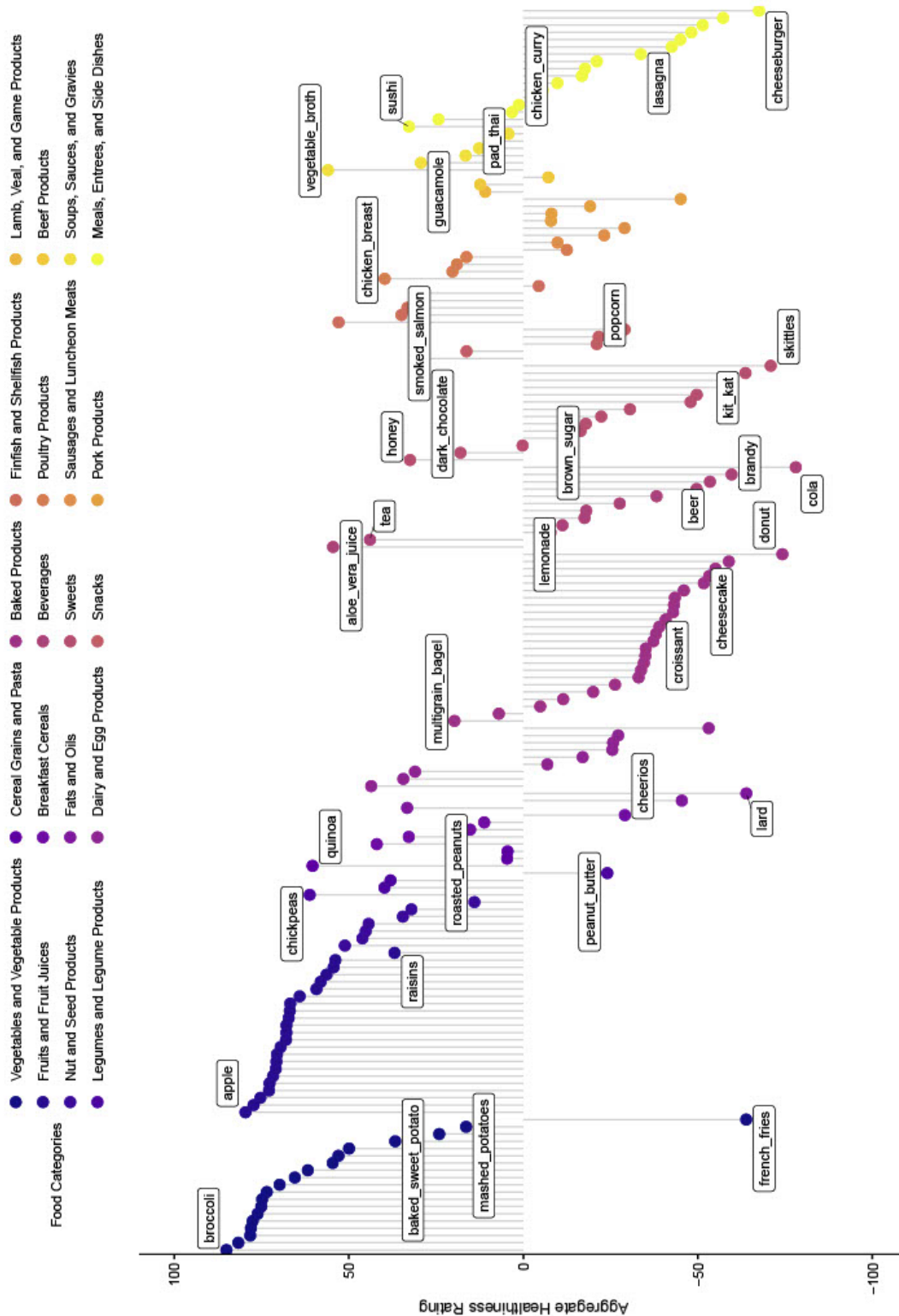
Finally, our third *Combined Model* concatenated the 11-dimensional Nutrient Model with the 300-dimensional Vector Representation Model. Using ridge regression, we explore the extent that both models can collectively explain people's subjective food healthiness judgments.

Results

We began by examining the distribution of aggregate healthiness ratings in Figure 1. Here we observed that healthiness judgments varied greatly amongst food stimuli, both across and within food categories. Unsurprisingly, the foods with the healthiest ratings were all fruit and vegetables, with the top five mean ratings ranging between 77 and 82 for tomatoes, cucumber, apple, carrots, and broccoli, respectively. The five foods that received the unhealthiest ratings, ranging between -65 and -50, were cola, donut, skittles, cheeseburger, and kit kat.

Figure 1

Distribution of aggregated food healthiness ratings from Study 1A



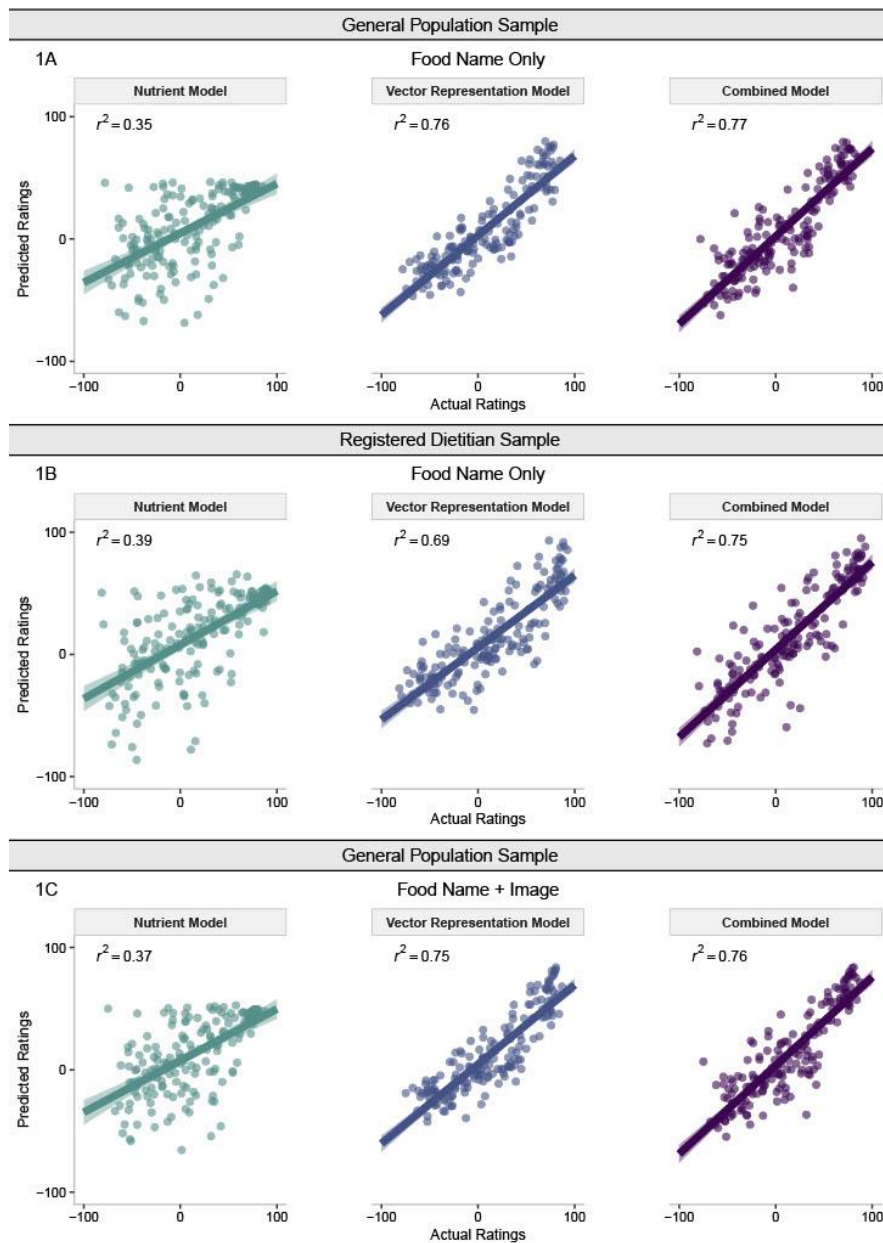
Note. For clarity, foods are separated by food category. Text labels indicate exemplar food items in each of the categories.

PREDICTING FOOD HEALTHINESS JUDGMENT

Figure 2 summarizes the accuracy rates of our three models in Studies 1A, 1B, and 1C. The dots within each scatterplot represent the out-of-sample predicted vs actual (aggregated) healthiness ratings for the individual foods. As we are using predictive modeling, the coefficient of determination (r^2) reflects the performance of the model when making out-of-sample predictions.

Figure 2

A Comparison of Predictive Accuracy between Models in Studies 1A, 1B, and 1C.



Note. A comparison of predictive accuracy between models that used only nutrient content, only word vector representations, or a combination of nutrient content and word vector representations in a general population sample (1A), expert sample (1B) and with food images included as stimuli (1C).

PREDICTING FOOD HEALTHINESS JUDGMENT

As shown in Figure 2, the out-of-sample predictive accuracy of the Vector Representation Model was very high across all studies, with an r^2 ranging from 0.69 (95% CI [0.63, 0.75]) to 0.76 (95% CI [0.71, 0.81]). By comparison, the predictive accuracy of the model based on the foods' nutritional information was always lower (r^2 ranging from 0.35, 95% CI [0.24, 0.46] to 0.39, 95% CI [0.28, 0.50]). The Combined Model performs best however, achieving marginally higher predictive accuracy than the Vector Representation Model in every study (r^2 ranging from 0.75 (95% CI [0.70, 0.80]) to 0.77, 95% CI [0.72, 0.82]). Overall, these findings highlight that the performance of the Vector Representation Model is stable, even when using ratings from participants with high nutritional expertise and with food images as stimuli.

We performed several robustness checks to assure the reliability of our findings. First, we ran separate paired sample t-tests to compare the squared errors from different models for each study (see Section 6 of the Supplementary Materials). Across all studies, the mean squared errors from the Vector Representation Model and the Combined Model were significantly lower than those from the Nutrient Model (all $p < 0.01$). We also repeated our analysis at the individual level, without aggregating healthiness ratings for each food. Results are presented in our Supplementary Materials (Section 7) and show that our findings remain largely unchanged. Section 4 of the Supplement summarizes r^2 for the Vector Representation Model based on alternative word vectors obtained from fastText (Mikolov et al., 2018) and GloVe (Pennington et al., 2014). Finally, in Section 5, we show the results of different regression techniques, including lasso, support vector, and k-nearest neighbors. Once again, using alternative word vectors or regression techniques did not alter our results.

Returning to the results from the Vector Representation Model based on the ridge regression and word2vec vectors, our approach was also able to capture qualitative trends in our data by correctly predicting the categories of foods judged as being high or low in healthiness. For example, both observed and predicted ratings were highest for categories such as Fruits and Fruit Juices, Vegetables and Vegetable Products, and Nut and Seed Products. Likewise, both observed and predicted ratings were lowest for categories such as Baked Products, Sweets, and Fats and Oils. In fact, when pooling the data by food category, we found the Vector Representation Model predicted average healthiness ratings for categories of foods with an out-of-sample r^2 of 0.83 (95% CI [0.79, 0.86]). The Nutrient Model, in contrast, achieved an r^2 of only 0.31 (95% CI [0.20, 0.41]). It seems healthiness judgments are sensitive to the category of the food item, a property easily captured by the

PREDICTING FOOD HEALTHINESS JUDGMENT

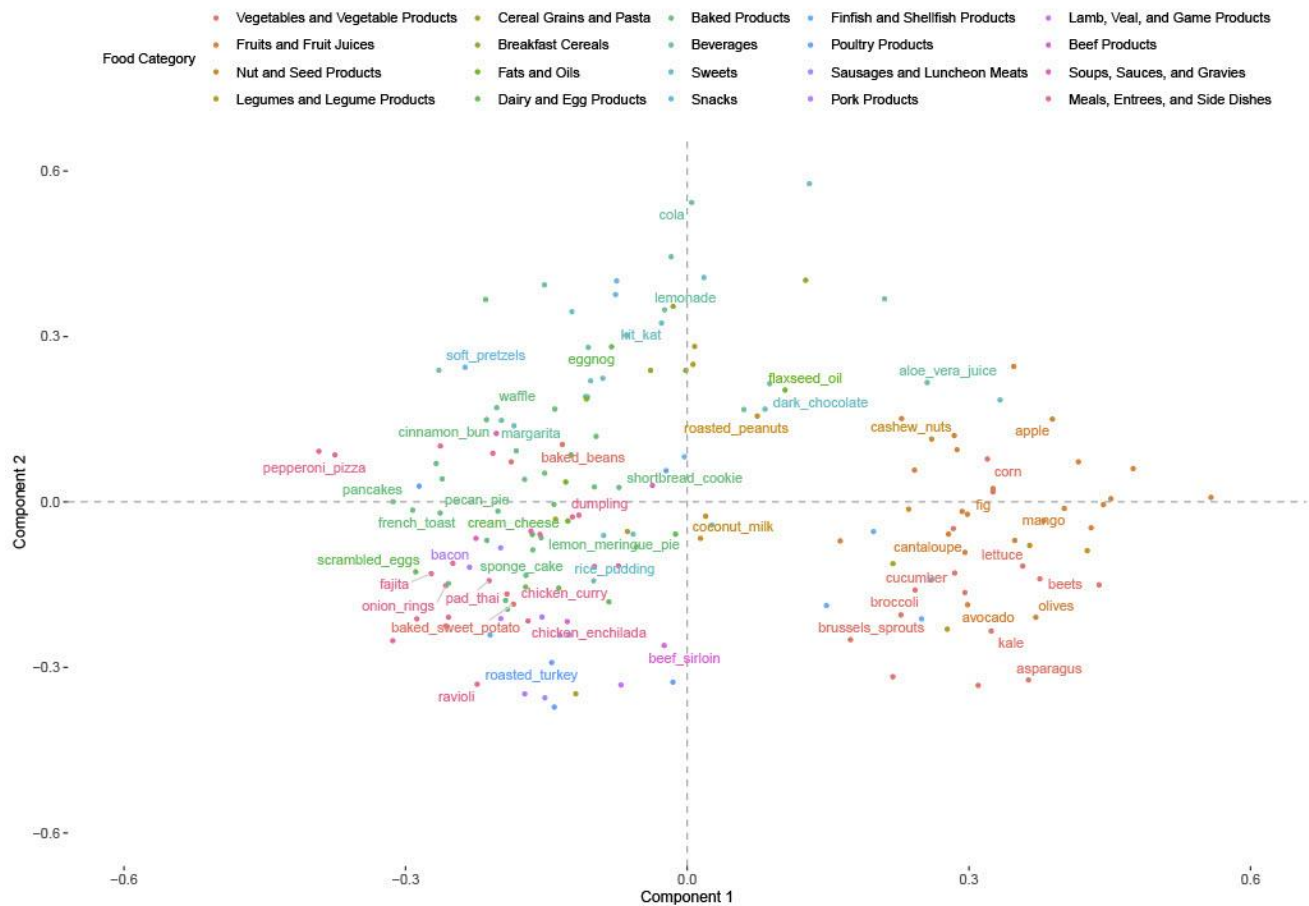
Vector Representation Model, but less so for the Nutrient Model (Orquin, 2014). Further details of this analysis are provided in Section 8 of the Supplementary Materials.

The reason why the Vector Representation Model performs well is that it may capture the latent associations underpinning judgments of healthiness. To explore these associations, we applied a Principal Component Analysis to the vector representations of the 172 food items. Projections for the first two components are shown in Figure 3. By inspecting Component 1, it is clear that negative values correspond to mostly heavily processed and junk foods (e.g., pepperoni pizza, bacon, onion rings), whereas positive values correspond mainly to organic and unprocessed vegetables and fruits (e.g. apple, mango, lettuce, beets). Component 2 on the other hand, appears to reflect the sweetness/sugariness of the food. The most positive scoring foods on this component are sugary drinks (e.g. cola, lemonade) and sugary snacks (e.g. kit kat, dark chocolate). Among the negative scores for Component 2, we can see meats (e.g., roasted turkey) but also less sugary vegetables (e.g., brussels sprouts, asparagus).

PREDICTING FOOD HEALTHINESS JUDGMENT

Figure 3

A 2d projection (based on the Principal Component Analysis) of vector representations for the 172 food names.



Note. For clarity, only a random subset of 50 names are labeled on the plot.

Another benefit of the vector representation approach is that it can identify regions of the semantic space related to food healthiness. This can be done by passing the vector representations of common words (that are not necessarily food items) through a model trained on participants' food healthiness judgments. Words given high predictions would be those most associated with healthiness and would capture the conceptual underpinnings of health judgment. Figure 4 shows a word cloud of the fifty English language words with the highest healthiness predictions, derived with this approach. Visibly, agriculture and nature-related words, such as crop, organic, and leaf, make up the majority of this word cloud. Interestingly, the word healthy is also present in the word cloud even though our model was never explicitly trained on this concept. It seems that implicit in people's judgments are associations with concepts like healthiness, as well as other concepts (e.g.,

PREDICTING FOOD HEALTHINESS JUDGMENT

rawness, which underpin food healthiness judgments. However, it is important to determine whether these associations continue to play a role even if foods' nutritional values are made more salient.

We addressed these issues in Studies 2A-2C by eliciting food healthiness judgments from participants who saw either food names alone (as in Study 1A) or food names along with the label highlighting various aspects of its nutrition. Again, we recruited adult participants for this series of studies. In Study 2A, we provided our treatment group with information about the calories per 100g. The provision of calorie information to aid healthy eating is supported by qualitative research showing that consumers use energy content information (calories) as a proxy for the overall nutritional value of a product (van Kleef et al., 2008). In Studies 2B and 2C, we examined the effects of information about key nutrients (fat, saturates, sugars, and salt). Under EU rules, front-of-pack labeling of this kind is acceptable with either no color-coding or traffic-light colored cues (i.e., red highlights high, orange medium, and green low amounts of fat, saturates, sugar, and salt) (European Commission, 2011). While both strategies highlight individual nutrients, it is the color-coded format that also aids consumers to judge whether a particular amount is high, medium, or low. In Study 2B we gave the treatment group nutrient labels without color-coding and in Study 2C we gave this group with nutrient labels with color-coding.

Methods

Participants

There were 202 participants in Study 2A, and after the removal of five using our exclusion criteria, 197 participants were included in the final analysis (aged 18-71 years, $M_{\text{age}} = 30.30$ years, $SD = 10.74$, 52% female, and 80% had no dietary restrictions). From the initial 199 participants who took part in Study 2B, four were excluded leaving 195 participants (aged 18-65 years, $M_{\text{age}} = 29.16$ years, $SD = 10.28$, 48% female, and 82% had no dietary restrictions). Finally, 202 participants took part in Study 2C (aged 18-78 years, $M_{\text{age}} = 34.69$ years, $SD = 11.51$, 70% female, and 81% had no dietary restrictions). No participants were excluded from this study as all participant responses fell above the threshold for removal. Only residents of the UK were allowed to participate in Study 2C to assure knowledge and familiarity with the traffic light food labeling system.

Design and Procedure

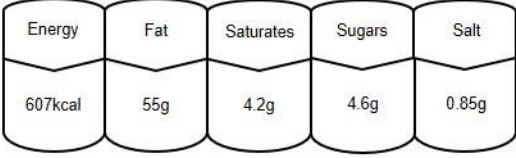
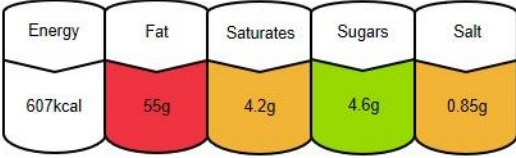
We tested the role of food labeling on judgment, in which we gradually (across studies) introduced more informative (and realistic) formats of food labeling. All three studies used a between-

PREDICTING FOOD HEALTHINESS JUDGMENT

subjects design. In half of the sample (control group) participants rated food healthiness of 172 foods in the same manner as in Study 1A and 1B. In the treatment groups, participants rated each of the food names presented alongside a nutrition label. In Study 2A, this was the energy (kcal) amounts per 100 grams of the food. In Study 2B, we additionally included the absolute amount of fats, saturates, sugars and salts. Finally, in Study 2C, we used the same objective information as above, but also added the “traffic light” system used in the UK, which indicates the relative amount of different nutrients, categorizing them into green, amber, and red groups. The examples of the labeling used in each study are presented in Figure 5.

Figure 5

Food stimuli presented to participants in Studies 2A, 2B and 2C

Study	Control Condition	Treatment Condition
2A	Food Name Only	Food Name + Typical values per 100g: Energy 607kcal
2B	Food Name Only	Food Name + Typical values per 100g of the food item are as follows: 
2C	Food Name Only	Food Name + Typical values per 100g of the food item are as follows: 

Note. All participants were from a general population sample.

Results

As shown in Figure 6, the Vector Representation Model performed very well across all studies and conditions. In fact, the out-of-sample predictive accuracy of the Vector Representation Model was very high, with r^2 ranging from 0.65 (95% CI [0.59, 0.72]) to 0.77 (95% CI [0.72, 0.81]), in each study and condition. By comparison, the predictive accuracy of the models based on the foods' nutritional information was lower but also much more variable (r^2 ranging from 0.33, 95% CI [0.22, 0.44] to 0.77,

PREDICTING FOOD HEALTHINESS JUDGMENT

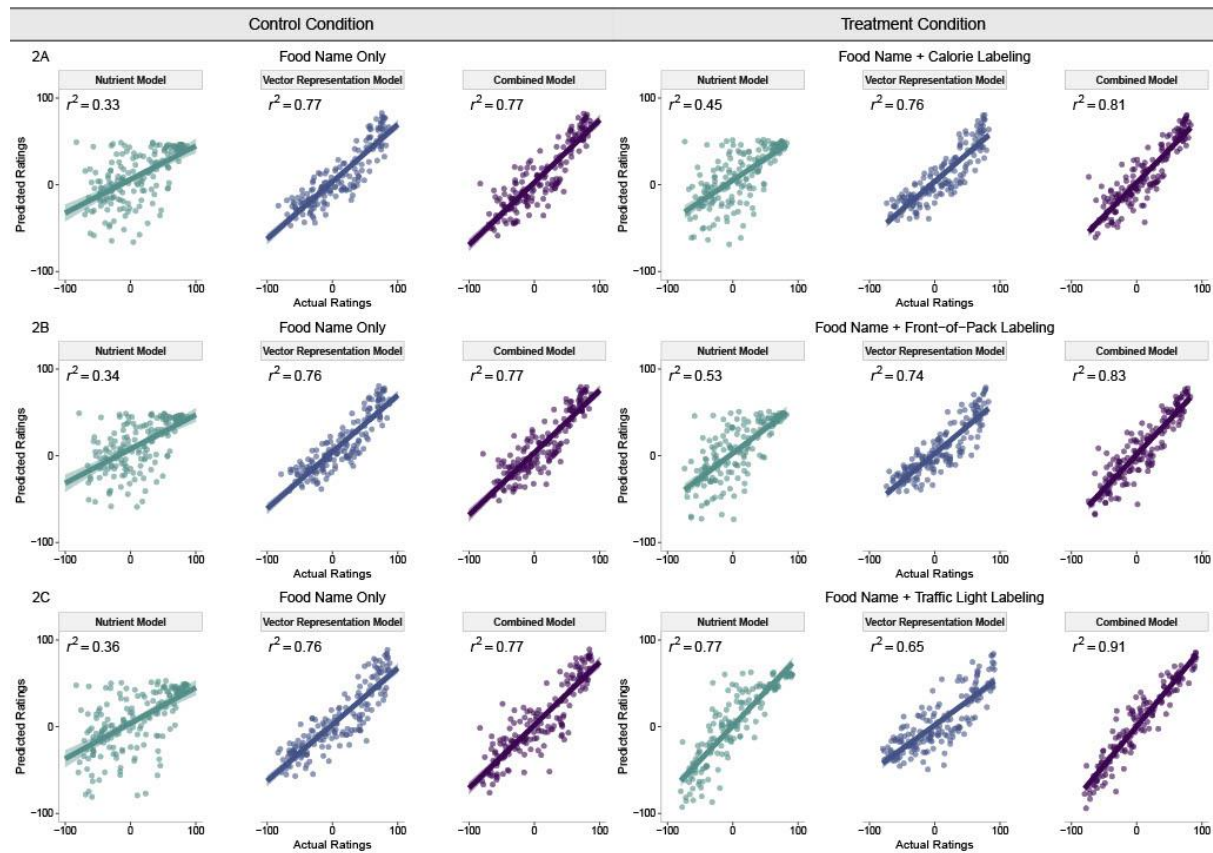
95% CI [0.72, 0.83]). Figure 6 reveals a systematic pattern – the predictive accuracy of the Nutrient Model increased with the amount of the nutritional information presented alongside foods' names. This is unsurprising as it shows that people integrated label information into their judgments (Gonzalez-Vallejo et al., 2016; Scarborough et al., 2007). Despite this, the Vector Representation Model still performed better than the Nutrient Model when participants saw only calorie information (Study 2A) and calorie information with front-of-pack nutrient labeling (Study 2B). Only in the most informative condition, traffic light labeling (Study 2C), did the Nutrient Model outperform the Vector Representation Model. Figure 6 also shows that the accuracy of the vector representation model is identical across the two conditions in Studies 2A and 2B, although it does drop slightly in Study 2C. This is not significant, as can be seen from the slight overlap in 95% CIs of the control ($r^2 = 0.76$, 95% CI = [0.71, 0.81]) and traffic light labelling conditions ($r^2 = 0.65$, 95% CI = [0.59, 0.72]) in Figure S3 of the Supplementary Materials. In any case, these results show that associations with food names play an important role in people's judgments of healthiness, often more than its nutritional composition.

Figure 6 also summarizes the predictive accuracy of the Combined Model – which uses both the word vectors *and* nutritional information to predict people's judgments. In 5 out of 6 cases, the Combined Model performed better than the individual models. The highest accuracy was achieved in the traffic light labeling condition, with r^2 of 0.91 (95% CI [0.89, 0.93]), which was markedly higher than 0.77 (95% CI [0.72, 0.83]) of the Nutrient Model and 0.65 (95% CI [0.59, 0.72]) of the Vector Representation Model. These results support the interpretation that word vectors explain people's judgments over and above the nutritional information of individual foods.

PREDICTING FOOD HEALTHINESS JUDGMENT

Figure 6

A Comparison of Predictive Accuracy between Models in Studies 2A, 2B and 2C



Note. A comparison of predictive accuracy between models that used only nutrient content, only word vector representations, or a combination of nutrient content and word vector representations. Actual ratings were all from a general population sample who were randomly assigned to either the control or the treatment condition in each study.

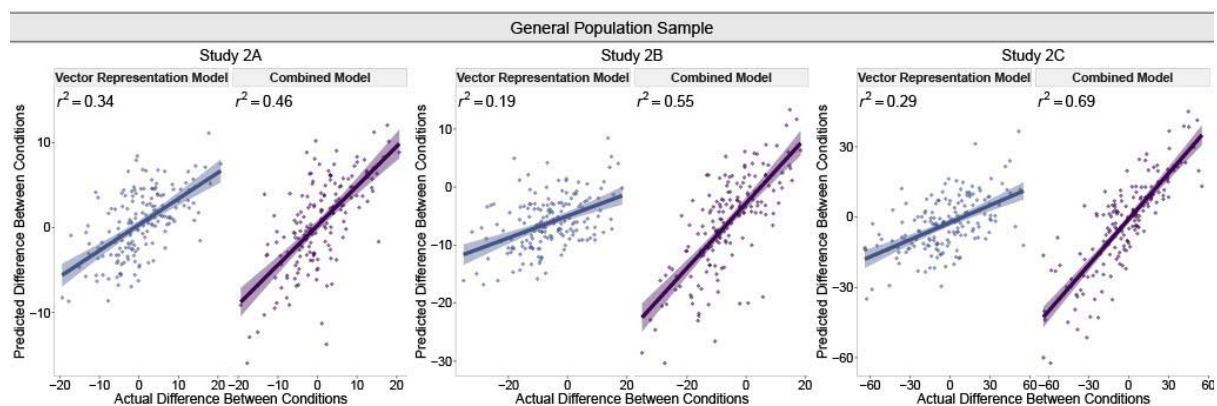
Do representations of foods change when nutrient information of foods is highlighted? In other words, do we observe a systematic shift in knowledge representations due to the various types of food labeling? To answer this question, we computed differences between aggregate ratings for each food made in the condition with and without a food label. We then refitted our Vector Representation and Combined Model with these difference scores as a dependent variable. Figure 7 shows that the Vector Representation Model explains a non-trivial amount of variance in the difference between the conditions in all three studies. At the same time, the predictive accuracy of the Combined Model increases markedly from Study 2A through to 2C, which confirms that people who saw additional

PREDICTING FOOD HEALTHINESS JUDGMENT

nutritional information did in fact rely on it when making their judgments. This supports the interpretation that even if food labeling changes how people make judgments, the reliance on knowledge representations captured by the Vector Representation Model remains stable and influential. It also shows that word vector representations can predict the idiosyncratic effects of nutrient labels on health judgments for different food items.

Figure 7

Leave-one-out cross-validation results for the ability of vector representations to predict condition differences



Note. Leave-one-out cross-validation results for the ability of vector representations (Vector Representation Model) and the Combined Model to predict the difference between conditions in Study 2A (calorie information – control), Study 2B (calorie information with front-of-pack nutrient labeling – control) and, Study 2C (traffic light labeling – control).

General Discussion

Everyday dietary decisions are influenced by people's subjective perceptions of food healthiness. Psychological explanations of this process are incomplete without an accurate model of the rich knowledge and diverse associations underpinning people's judgments of what food is healthy and what is not. In this paper, we offer a novel method for uncovering these knowledge representations by combining insights from machine learning and computational linguistics. Using vector representations of food items derived from natural language, we show that it is possible to predict healthiness judgments highly accurately. We show that people's judgments can be partly explained by the strength of association between individual food items and concepts pertaining to

PREDICTING FOOD HEALTHINESS JUDGMENT

naturalness (e.g., harvest, leaf) and rawness (e.g., crop, organic). These associations play a role even when judgments are made in the presence of food images or are made by trained dietitians. In addition, high accuracy rates obtained by our Combined Model indicate that such knowledge representations in language do not merely reflect beliefs about nutritional composition; rather they capture something unique about people's associations with different foods. Thus, our models can help evaluate how different front-of-pack labeling strategies influence food healthiness judgments.

Unlike previous approaches, our method does not require us to identify specific factors or attributes that we, as researchers, believe to be related to healthiness judgments. Instead, by using our best-fit model to predict the "healthiness" of common words in the English language, we show that nature-related words such as "crop", "harvest", and "agricultural" are implicit in people's judgments. This is consistent with other findings that perceived naturalness and healthiness are often intertwined (Sanchez-Siles et al., 2019; Siipi, 2012 but see Fernbach et al., 2019). These results also align with the finding that rawness or the degree to which a food has been processed is a strong cue of healthiness in food choice (Scheibehenne et al., 2007; Schulte-Mecklenbeck et al., 2013). Notably, our results indicate that models based on these associations are accurate even if participants are explicitly told about the nutritional composition of foods.

Our approach offers a unique insight into the psychological basis of subjective food healthiness judgments by exploring foods in their most abstract forms (name or image). That said, a model trained on written text is unlikely to accurately capture sensorimotor information about foods (e.g. smell, texture), which would also be relevant in real-world situations (De Deyne et al., 2016; Lynott et al., 2020; Papiés et al., 2020). Hence, while our results are promising, they are only a first step in providing a rich set of attributes and associations that people use in judging foods' healthiness.

Neither explicit food labeling nor expert judgments reduced the contributions of the knowledge associations established by the Vector Representation model. With respect to expert judgments, these findings are in line with research showing that nutritional expertise does not always translate into a higher reliance on nutritional information when making healthiness judgments (Orquin, 2014). Our results also speak to the value of nutritional labeling more generally. Given that associations played a role in all studies, existing front-of-pack labeling can neither substitute nor correct for the associations that people rely on when judging foods' healthiness.

PREDICTING FOOD HEALTHINESS JUDGMENT

There are many potentially useful applications of our computational approach. Future studies could test the predictive ability of this Vector Representation Model with and against other formats of nutrient labeling such as France's Nutri-score label (color-coded without numerical information). Thus, the use of this approach could be vital in determining a single internationally agreed nutrient labeling system (Goiana-da-Silva et al., 2019), especially since it provides directly comparable results between labeling formats. However, further work is necessary to establish whether the accuracy of our models changes when participants are presented with other information present on pre-packaged foods, such as branding, health claims, and back-of-pack nutrition labeling.

An important outstanding question is whether our Vector Representation Model is generalizable to judgments of other foods than the 172 items tested in all six studies. In Section 10 of our Supplementary Material, we report the results of a new study in which we elicited judgments of 60 new foods from a sample of 97 participants. Instead of training a new model, we used the Vector Representation Model from Study 1A to derive predictions for our new foods. Our models performed very well –with our approach we can predict healthiness judgments of new foods from a new group of participants highly accurately. To assist future research, we have obtained predictions of our models for hundreds of novel food items and made these available via OSF (<https://osf.io/jys6u/>). These can be used to evaluate future interventions and to test alternative psychological mechanisms that underpin human judgments and choices of foods. Overall, our studies provide new insights into people's food healthiness judgments, while our methods offer an exciting new avenue to researchers and practitioners interested in designing interventions for healthy eating.

Open Practices Statement

The studies reported in this article were not preregistered. De-identified data and code to replicate the findings are available via the Open Science Framework (<https://osf.io/jys6u/>).

Acknowledgments

This work was supported by the Engineering and Physical Sciences Research Council (grant EP/N509796/1: project reference no.1939178) and the National Science Foundation (grant SES-1847794).

References

- Bauer, U. E., Briss, P. A., Goodman, R. A., & Bowman, B. A. (2014). Prevention of chronic disease in the 21st century: elimination of the leading preventable causes of premature death and disability in the USA. *The Lancet*, *384*(9937), 45-52.
[https://doi.org/https://doi.org/10.1016/S0140-6736\(14\)60648-6](https://doi.org/https://doi.org/10.1016/S0140-6736(14)60648-6)
- Bhatia, S. (2017). Associative judgment and vector space semantics. *Psychological Review*, *124*(1), 1-20. <https://doi.org/https://doi.org/10.1037/rev0000047>
- Bhatia, S. (2019). Predicting Risk Perception: New Insights from Data Science. *Management Science*, *65*(8), 3800-3823. <https://doi.org/https://doi.org/10.1287/mnsc.2018.3121>
- Blechert, J., Lender, A., Polk, S., Busch, N., & Ohla, K. (2019). Food-pics_extended—an image database for experimental research on eating and appetite: additional images, normative ratings and an updated review. *Frontiers in Psychology*, *10*, 307.
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6416180/pdf/fpsyg-10-00307.pdf>
- Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183-186.
<https://doi.org/https://doi.org/10.1126/science.aal4230>
- De Deyne, S., Perfors, A., & Navarro, D. J. (2016). Predicting human similarity judgments with distributional models: The value of word associations. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 1861-1870.
<https://doi.org/https://doi.org/10.1126/science.aal4230>
- Downs, J. S., Wisdom, J., & Loewenstein, G. (2015). Helping consumers use nutrition information: Effects of format and presentation. *American Journal of Health Economics*, *1*(3), 326-344.
https://doi.org/10.1162/AJHE_a_00020
- European Commission. (2011). Regulation (EU) No 1169/2011 of the European Parliament and of the Council of 25 October 2011 on the provision of food information to consumers, amending Regulations (EC) No 1924/2006 and (EC) No 1925/2006 of the European Parliament and of the Council, and repealing Commission Directive 87/250/EEC, Council Directive 90/496/EEC, Commission Directive 1999/10/EC, Directive 2000/13/EC of the European Parliament and of

PREDICTING FOOD HEALTHINESS JUDGMENT

- the Council, Commission Directives 2002/67/EC and 2008/5/EC and Commission Regulation (EC) No 608/2004. *Official Journal of the European Union*, 304, 18-63.
- Fernbach, P. M., Light, N., Scott, S. E., Inbar, Y., & Rozin, P. (2019). Extreme opponents of genetically modified foods know the least but think they know the most. *Nature Human Behaviour*, 3(3), 251. <https://doi.org/https://doi.org/10.1038/s41562-018-0520-3>
- Goiana-da-Silva, F., Cruz-e-Silva, D., Miraldo, M., Calhau, C., Bento, A., Cruz, D., Almeida, F., Darzi, A., & Araújo, F. (2019). Front-of-pack labelling policies and the need for guidance. *The Lancet Public Health*, 4(1). [https://doi.org/https://doi.org/10.1016/S2468-2667\(18\)30256-1](https://doi.org/https://doi.org/10.1016/S2468-2667(18)30256-1)
- Kanter, R., Vanderlee, L., & Vandevijvere, S. (2018). Front-of-package nutrition labelling policy: global progress and future directions. *Public Health Nutrition*, 21(8), 1399-1408. <https://doi.org/https://doi.org/10.1017/S1368980018000010>
- Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2), 211-240. <https://doi.org/https://doi.org/10.1037/0033-295X.104.2.211>
- Lenci, A. (2018). Distributional Models of Word Meaning. *Annual Review of Linguistics*, 4(1), 151-171. <https://doi.org/https://doi.org/10.1146/annurev-linguistics-030514-125254>
- Lobstein, T., & Davies, S. (2008). Defining and labelling 'healthy' and 'unhealthy' food. *Public Health Nutrition*, 12(3), 331-340. <https://doi.org/https://doi.org/10.1017/S1368980008002541>
- Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2020). The Lancaster Sensorimotor Norms: multidimensional measures of perceptual and action strength for 40,000 English words. *Behav Res Methods*, 52(3), 1271-1291. <https://doi.org/10.3758/s13428-019-01316-z>
- Mikolov, T., Grave, E., Bojanowski, P., Puhersch, C., & Joulin, A. (2018). Advances in pre-training distributed word representations. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *NeurIPS*, 3111-3119.
- Orquin, J. L. (2014). A Brunswik lens model of consumer health judgments of packaged foods. *Journal of Consumer Behaviour*, 13(4), 270-281. <https://doi.org/https://doi.org/10.1002/cb.1465>

PREDICTING FOOD HEALTHINESS JUDGMENT

- Papies, E. K. (2013). Tempting food words activate eating simulations. *Front Psychol*, 4, 838.
<https://doi.org/10.3389/fpsyg.2013.00838>
- Papies, E. K., Barsalou, L. W., & Rusz, D. (2020). Understanding Desire for Food and Drink: A Grounded-Cognition Approach. *Current Directions in Psychological Science*, 29(2), 193-198.
<https://doi.org/10.1177/0963721420904958>
- Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532-1543.
- Pereira, F., Gershman, S., Ritter, S., & Botvinick, M. (2016). A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cognitive Neuropsychology*, 33(3-4), 175-190.
<https://doi.org/https://doi.org/10.1080/02643294.2016.1176907>
- Perkovic, S., & Orquin, J. L. (2018). Implicit Statistical Learning in Real-World Environments Leads to Ecologically Rational Decision Making. *Psychol Sci*, 29(1), 34-44.
<https://doi.org/10.1177/0956797617733831>
- Pieniak, Z., Verbeke, W., Vanhonacker, F., Guerrero, L., & Hersleth, M. (2009). Association between traditional food consumption and motives for food choice in six European countries. *Appetite*, 53(1), 101-108. <https://doi.org/https://doi.org/10.1016/j.appet.2009.05.019>
- Plasek, B., Lakner, Z., & Temesi, Á. (2020). Factors that Influence the Perceived Healthiness of Food. *Nutrients*, 12(6), 1881.
- Reutner, L., Genschow, O., & Wänke, M. (2015). The adaptive eater: Perceived healthiness moderates the effect of the color red on consumption. *Food Quality and Preference*, 44, 172-178.
- Richie, R., Zou, W., & Bhatia, S. (2019). Predicting High-Level Human Judgment Across Diverse Behavioral domains. *Collabra: Psychology*, 5(1), 50.
<https://doi.org/http://doi.org/10.1525/collabra.282>
- Sanchez-Siles, L. M., Michel, F., Román, S., Bernal, M. J., Philipsen, B., Haro, J. F., Bodenstab, S., & Siegrist, M. (2019). The Food Naturalness Index (FNI): An integrative tool to measure the degree of food naturalness. *Trends in Food Science & Technology*, 91, 681-690.
<https://doi.org/10.1016/j.tifs.2019.07.015>

PREDICTING FOOD HEALTHINESS JUDGMENT

- Sanjari, S. S., Jahn, S., & Boztug, Y. (2017). Dual-process theory and consumer response to front-of-package nutrition label formats. *Nutr Rev*, 75(11), 871-882.
<https://doi.org/https://doi.org/10.1093/nutrit/nux043>
- Scarborough, P., Boxer, A., Rayner, M., & Stockley, L. (2007). Testing nutrient profile models using data from a survey of nutrition professionals. *Public Health Nutrition*, 10(4), 337-345.
<https://doi.org/https://doi.org/10.1017/S1368980007666671>
- Scheibehenne, B., Miesler, L., & Todd, P. M. (2007). Fast and frugal food choices: Uncovering individual decision heuristics. *Appetite*, 49(3), 578-589.
<https://doi.org/https://doi.org/10.1016/j.appet.2007.03.224>
- Schuldt, J. P., & Schwarz, N. (2010). The "organic" path to obesity? Organic claims influence calorie judgments and exercise recommendations. *Judgment and Decision Making*, 5(3), 144-150.
- Schulte-Mecklenbeck, M., Sohn, M., de Bellis, E., Martin, N., & Hertwig, R. (2013). A lack of appetite for information and computation. Simple heuristics in food choice. *Appetite*, 71, 242-251.
- Siipi, H. (2012). Is Natural Food Healthy? *Journal of Agricultural and Environmental Ethics*, 26(4), 797-812. <https://doi.org/10.1007/s10806-012-9406-y>
- Soederberg Miller, L. M., & Cassady, D. L. (2015). The effects of nutrition knowledge on food label use. A review of the literature. *Appetite*, 92, 207-216.
<https://doi.org/https://doi.org/10.1016/j.appet.2015.05.029>
- Step toe, A., Pollard, T. M., & Wardle, J. (1995). Development of a measure of the motives underlying the selection of food: the food choice questionnaire. *Appetite*, 25(3), 267-284.
- Trudel, R., Murray, K. B., Kim, S., & Chen, S. (2015). The impact of traffic light color-coding on food health perceptions and choice. *J Exp Psychol Appl*, 21(3), 255-275.
<https://doi.org/10.1037/xap0000049>
- Turnwald, B. P., Boles, D. Z., & Crum, A. J. (2017). Association between indulgent descriptions and vegetable consumption: twisted carrots and dynamite beets. *JAMA internal medicine*, 177(8), 1216-1218.
- U.S. Department of Agriculture, A. R. S. (2019). *FoodData Central*.
<https://doi.org/https://fdc.nal.usda.gov/index.html>

PREDICTING FOOD HEALTHINESS JUDGMENT

van Kleef, E., van Trijp, H., Paeps, F., & Fernandez-Celemin, L. (2008). Consumer preferences for front-of-pack calories labelling. *Public Health Nutr*, 11(2), 203-213.

<https://doi.org/10.1017/S1368980007000304>

Whalen, R., Harrold, J. A., Child, S., Halford, J. C., & Boyland, E. J. (2018). The Health Halo Trend in UK Television Food Advertising Viewed by Children: The Rise of Implicit and Explicit Health Messaging in the Promotion of Unhealthy Foods. *Int J Environ Res Public Health*, 15(3), 560.

<https://doi.org/https://doi.org/10.3390/ijerph15030560>

Zou, W., & Bhatia, S. (2021). Judgment errors in naturalistic numerical estimation. *Cognition*, 211, 104647.

Supplementary Materials

Table of Contents

1. Full Sample Results
2. Demographic Characteristics
3. Extended Nutrient and Combined Models
4. Alternative Word Embeddings
5. Secondary Vector Representation Models
6. Test for Model Comparison
7. Individual-level Modeling
8. Food Category Modeling
9. Word Cloud for Unhealthy Associations
10. Test of Model Generalizability
11. References

1. Full Sample Results

In our main manuscript, we excluded participants whose ratings correlated poorly ($r < 0.4$) with average participant ratings. To make sure that the exclusion criterion did not significantly alter the predictive accuracy of our models, we also fit the three main models to the original data without the exclusion. Table S1 reports the out-of-sample r^2 with either 95% CI for the Nutrient Model, Vector Representation Model, and Combined Model trained on average healthiness ratings. Compared to the results reported in Table S6, model performance stayed the same when including all participants.

Table S1

Aggregate level out-of-sample r^2 for the Vector Representation Model, Nutrient Model and Combined Model fit to the original data including all participants, numbers in the brackets represent 95% CIs

Study	Nutrient Model	Vector Representation Model	Combined Model
1A	0.35 [0.24, 0.46]	0.76 [0.71, 0.80]	0.76 [0.71, 0.82]
1B	0.39 [0.28, 0.50]	0.69 [0.63, 0.75]	0.75 [0.70, 0.80]
1C	0.37 [0.26, 0.48]	0.76 [0.71, 0.80]	0.77 [0.72, 0.82]
2A_control	0.33 [0.23, 0.44]	0.77 [0.72, 0.81]	0.77 [0.72, 0.82]
2A_treatment	0.45 [0.35, 0.56]	0.76 [0.71, 0.81]	0.81 [0.77, 0.85]
2B_control	0.34 [0.23, 0.45]	0.76 [0.71, 0.81]	0.77 [0.72, 0.82]
2B_treatment	0.53 [0.44, 0.63]	0.74 [0.69, 0.79]	0.82 [0.79, 0.86]
2C_control	0.36 [0.25, 0.47]	0.76 [0.71, 0.81]	0.77 [0.72, 0.82]
2C_treatment	0.77 [0.72, 0.83]	0.65 [0.59, 0.72]	0.91 [0.89, 0.93]

2. Demographic Characteristics

Table S2

Demographic Characteristics of all Participants

Characteristic	Study 1A		Study 1B		Study 1C		Study 2A				Study 2B				Study 2C				Supplement	
	Control		Control		Food Images		Control		Calorie Labeling		Control		Front of Pack Labeling		Control		Traffic Light Labeling		New food stimuli	
	(Lay)		(Expert)		(Lay)		(Lay)		(Lay)		(Lay)		(Lay)		(Lay)		(Lay)		(Lay)	
Age																				
<i>N</i>	134		19		99		96		101		104		91		102		100		97	
Mean ± SD	29.57 ± 8.86		35.84 ± 10.36		27.25 ± 10.20		31.64 ± 11.66		28.96 ± 9.81		29.97 ± 11.27		28.35 ± 9.28		35.18 ± 10.60		34.19 ± 12.41		25.71 ± 8.97	
Min - Max	18 - 74		23 - 56		18 - 69		18 - 71		18 - 60		18 - 65		18 - 57		18 - 64		18 - 78		18 - 60	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
Gender																				
Female	58	43	17	89	44	44	50	52	53	52	42	40	51	56	77	75	64	64	38	39
Male	73	54	2	11	54	55	45	47	47	47	62	60	38	42	25	25	36	36	58	60
Other	2	1	0	0	1	1	1	1	0	0	0	0	1	1	0	0	0	0	1	1
Prefer not to say	1	1	0	0	0	0	0	0	1	1	0	0	1	1	0	0	0	0	0	0
Diet Restrictions																				
None	113	84	13	68	81	82	71	74	86	85	85	82	75	82	82	80	82	82	85	88
Pescetarian	1	1	2	11	3	3	7	7	4	4	2	2	4	4	4	4	3	3	2	2
Vegetarian	9	7	1	5	6	6	3	3	3	3	4	4	4	4	5	5	4	4	5	5
Vegan	2	1	0	0	2	2	3	3	1	1	4	4	2	2	4	4	5	5	2	2
Other	9	7	3	16	7	7	12	13	7	7	9	9	6	7	7	7	6	6	3	3

Table S3*Demographic Characteristics specific to Nutritional Experts*

Characteristic	Study 1B	
	Control (Expert)	
	<i>n</i>	%
No. of Years as a Registered Dietitian		
Less than 1 year	1	5
1-4 years	8	42
5-9 years	3	16
10-19 years	3	16
20 years or more	4	21
Area of Specialism		
Diabetes	3	16
Gastroenterology	4	21
Older People	1	5
Oncology	1	5
Pediatric	2	11
Parenteral and Enteral Nutrition	1	5
Renal Nutrition	6	32
Other	1	5

3. Extended Nutrient and Combined Models

In testing the accuracy of the Nutrient Model our assumption was that participants' healthiness ratings may be best reflected in people's knowledge of the nutritional/energy composition of individual foods. In our main manuscript, our Nutrient Model included information about calories and nutrients (fat, saturates, sugar, salt, and protein) per 100g, as well as the traffic light color-coding. Whereas this approach reflects current EU regulation about food labeling, it may not truly reflect the nutritional information that people actually rely on when making their judgments. In the following section, we provide further robustness checks to the results we report in the main manuscript by using three different extended versions of the Nutrient Model. First, most of the existing regulation is focused on highlighting nutrients that are typically associated with weight gain and poorer health. We, therefore, extended our original Nutrient Model by adding (up to) 23 positive nutrient variables (including fiber, calcium, and vitamin C). Second, we consider the possibility that perceptions of healthiness may better align with nutritional information expressed in relation to the portion size amounts of each food.

As there is no agreement of what portion size should be, we made a judgment call to use nutrient amounts per 100 calories (thus g/100kcal). This way, our Nutrient Model does not punish foods that would never be eaten in large volumes (e.g., chewing gum).

Since many of the foods do not have all the micronutrients, we used ridge regression instead of linear (ordinary least square) regression for the two alternative Nutrient Models and Combined Models with micronutrients added. Further analysis also confirmed that ridge regression is more appropriate for the extended Nutrient Model as it achieved higher predictive accuracy than linear regression.

As can be seen in Tables S4 and S5, the use of portion size amounts does not improve the accuracy of either the Nutrient Model or the Combined Model in any condition or study. The addition of positive nutrients does not improve the accuracy of the Combined Model but impairs the accuracy of the Nutrient Model. The extended Nutrient Model with the added micronutrients (either per 100g or per 100kcal) performed considerably worse than the original Nutrient Model. The predictive power of the extended Nutrient Model was highly variable (r^2 ranging between 0.07 (95% CI [0.01, 0.14]) to 0.72 (95% CI [0.66, 0.78])), increasing across the conditions where nutritional information was accessible to participants and performed best in the traffic light labeling condition of Study 2C. Similarly, the alternative Combined Models performed consistently with the original Combined Model. In all control conditions and Studies 1A and 1C, the r^2 of the extended Combined Model with micronutrients (either per 100g or per 100kcal) varied between 0.65 (95% CI [0.57, 0.73]) and 0.89 (95% CI [0.86, 0.92]). It performed slightly worse in the sample of registered dietitians and reached the highest out-of-sample accuracy in the treatment condition of Study 2C. Moreover, using separate paired-sample t-tests for each study, we found that the mean squared errors from the extended versions of the Nutrient Model were not significantly different from those of the original Nutrient Model. This result holds for the Combined Model, suggesting that adding more nuanced nutrient information did not improve the predictive accuracy of the Nutrient Model or the Combined Model. The detailed statistics are available upon request. Taken together, these robustness checks show that the relatively lower accuracy of the Nutrient Model was not due to the incorrect assumption about the role of micronutrients or portion size considerations in people's judgments.

Table S4*Out-of-sample r^2 comparisons between the original and alternative Nutrient Models.*

Study	Nutrient Model (per 100g)	Nutrient Model (per 100 kcal)	Nutrient Model with micronutrients (per 100g)	Nutrient Model with micronutrients (per 100kcal)
1A	0.35 [0.24, 0.46]	0.34 [0.23, 0.45]	0.12 [0.05, 0.20]	0.12 [0.05, 0.19]
1B	0.39 [0.28, 0.50]	0.38 [0.27, 0.49]	0.15 [0.07, 0.23]	0.14 [0.06, 0.22]
1C	0.37 [0.26, 0.48]	0.33 [0.23, 0.44]	0.18 [0.09, 0.26]	0.11 [0.04, 0.18]
2A_control	0.33 [0.22, 0.44]	0.32 [0.21, 0.43]	0.08 [0.02, 0.14]	0.07 [0.01, 0.14]
2A_treatment	0.45 [0.35, 0.56]	0.42 [0.31, 0.52]	0.26 [0.17, 0.35]	0.24 [0.15, 0.33]
2B_control	0.34 [0.23, 0.45]	0.32 [0.21, 0.43]	0.08 [0.02, 0.14]	0.11 [0.04, 0.19]
2B_treatment	0.53 [0.44, 0.63]	0.49 [0.39, 0.59]	0.37 [0.28, 0.46]	0.35 [0.26, 0.16]
2C_control	0.36 [0.25, 0.47]	0.35 [0.24, 0.46]	0.11 [0.04, 0.18]	0.09 [0.03, 0.16]
2C_treatment	0.77 [0.72, 0.83]	0.76 [0.70, 0.82]	0.74 [0.68, 0.79]	0.72 [0.66, 0.78]

Table S5*Out-of-sample r^2 comparisons between the original and alternative Combined Models.*

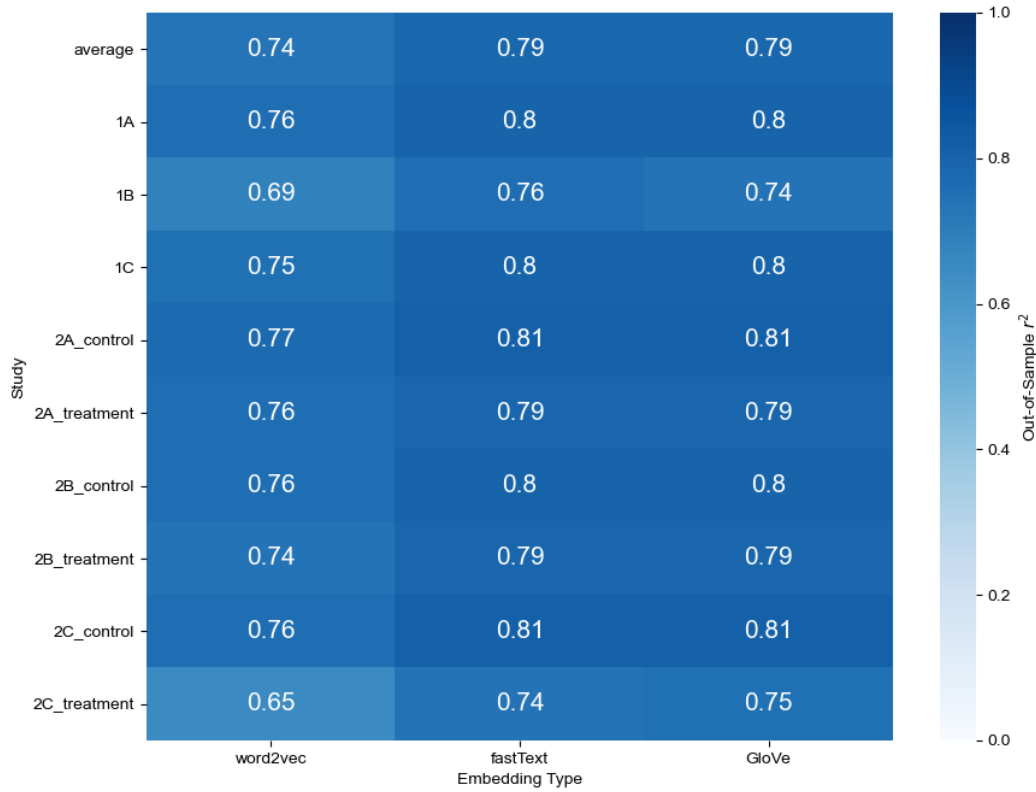
Study	Combined Model (per 100g)	Combined Model (per 100 kcal)	Combined Model with micronutrients (per 100g)	Combined Model with micronutrients (per 100kcal)
1A	0.77 [0.72, 0.82]	0.76 [0.71, 0.81]	0.66 [0.58, 0.74]	0.74 [0.67, 0.80]
1B	0.75 [0.70, 0.80]	0.72 [0.66, 0.77]	0.59 [0.50, 0.68]	0.67 [0.60, 0.75]
1C	0.76 [0.71, 0.82]	0.75 [0.70, 0.80]	0.68 [0.60, 0.75]	0.73 [0.66, 0.79]
2A_control	0.77 [0.72, 0.82]	0.77 [0.72, 0.82]	0.66 [0.58, 0.74]	0.73 [0.67, 0.80]
2A_treatment	0.81 [0.77, 0.85]	0.79 [0.74, 0.83]	0.73 [0.66, 0.80]	0.77 [0.71, 0.82]
2B_control	0.77 [0.72, 0.82]	0.76 [0.71, 0.81]	0.66 [0.58, 0.74]	0.73 [0.67, 0.80]
2B_treatment	0.83 [0.79, 0.86]	0.81 [0.76, 0.85]	0.75 [0.69, 0.81]	0.79 [0.73, 0.84]
2C_control	0.77 [0.72, 0.82]	0.77 [0.72, 0.82]	0.65 [0.57, 0.73]	0.74 [0.68, 0.80]
2C_treatment	0.91 [0.89, 0.93]	0.90 [0.80, 0.92]	0.88 [0.85, 0.91]	0.89 [0.86, 0.92]

4. Alternative Word Embeddings

In the main text, we fit the Vector Representation Model using pre-trained word2vec embeddings (Mikolov, et al., 2013) because these contained embeddings for multi-word food items. To show the robustness of using word embeddings in the Vector Representation Model, we also used other pre-trained word embeddings including fastText (Mikolov et al., 2018) and GloVe (Pennington et al., 2014), which both offer 300-dimensional vector representations of the food items. However, only 112 out of 172 food items have vector representation in fastText and 111 in GloVe. Unlike word2vec, which was trained on a large dataset of Google News articles, the GloVe model was trained on the Common Crawl 840B corpus and fastText on the Common Crawl 600B corpus. These embeddings also differ in terms of training algorithms. The pre-trained word2vec that we used was trained by a combination of the continuous bag-of-words (CBOW) method (which predicts words from their neighbors) and the skip-gram method (which predicts neighboring words of a given word). GloVe was trained in such a way that the dot product between two word vectors approximates the logarithm of the two words' probability of co-occurrence. FastText combines the CBOW method and position-dependent weighting to learn n-gram embeddings, which are summed to create word-level embeddings. Figure S1 shows out-of-sample r^2 of the Vector Representation Model (using ridge regression with $\lambda=1$) with different semantic vector representations from different word embeddings for every study. As can be seen, the performance of the Vector Representation Model with fastText or GloVe is comparable and even higher than word2vec. We suspect the higher performance with fastText and GloVe is due to the fact that the training corpus (Common Crawl) underpinning these two pre-trained models contains a more diverse set of language data than that of word2vec (Google News).

Figure S1

Out of Sample Accuracy from Alternative Word Embedding Sources



Note. Out-of-sample r^2 of the Vector Representation Models (VRM) with different semantic vector representations from alternative word embeddings. The same set of 172 food items were used in VRM with word2vec. However, only 112 were used in the Vector Representation Model with fastText and 111 in GloVe due to the limited vocabulary of these embeddings.

5. Secondary Vector Representation Models

In order to address the issue of having a high number of potentially highly correlated predictors in the Vector Representation Model (and Combined Model), we used a ridge regression. To check whether our results are influenced by the choice of method, we tested other regression techniques including lasso, support vector, and k-nearest neighbors regression. Here we provide brief details about these techniques:

Both ridge and lasso learn a linear mapping from semantic vector representations (x_{ij}) to healthiness ratings (y_i) while penalizing the magnitudes of coefficients (β_j) to avoid potential collinearities between dimensions (j). However, they differ in the loss function used for estimating

coefficients. Ridge penalizes the L2-norm of the coefficients (Eq. 1), whereas lasso penalizes the L1-norm of these coefficients (Eq. 2). Penalization parameter, λ , controls the strength of the penalty.

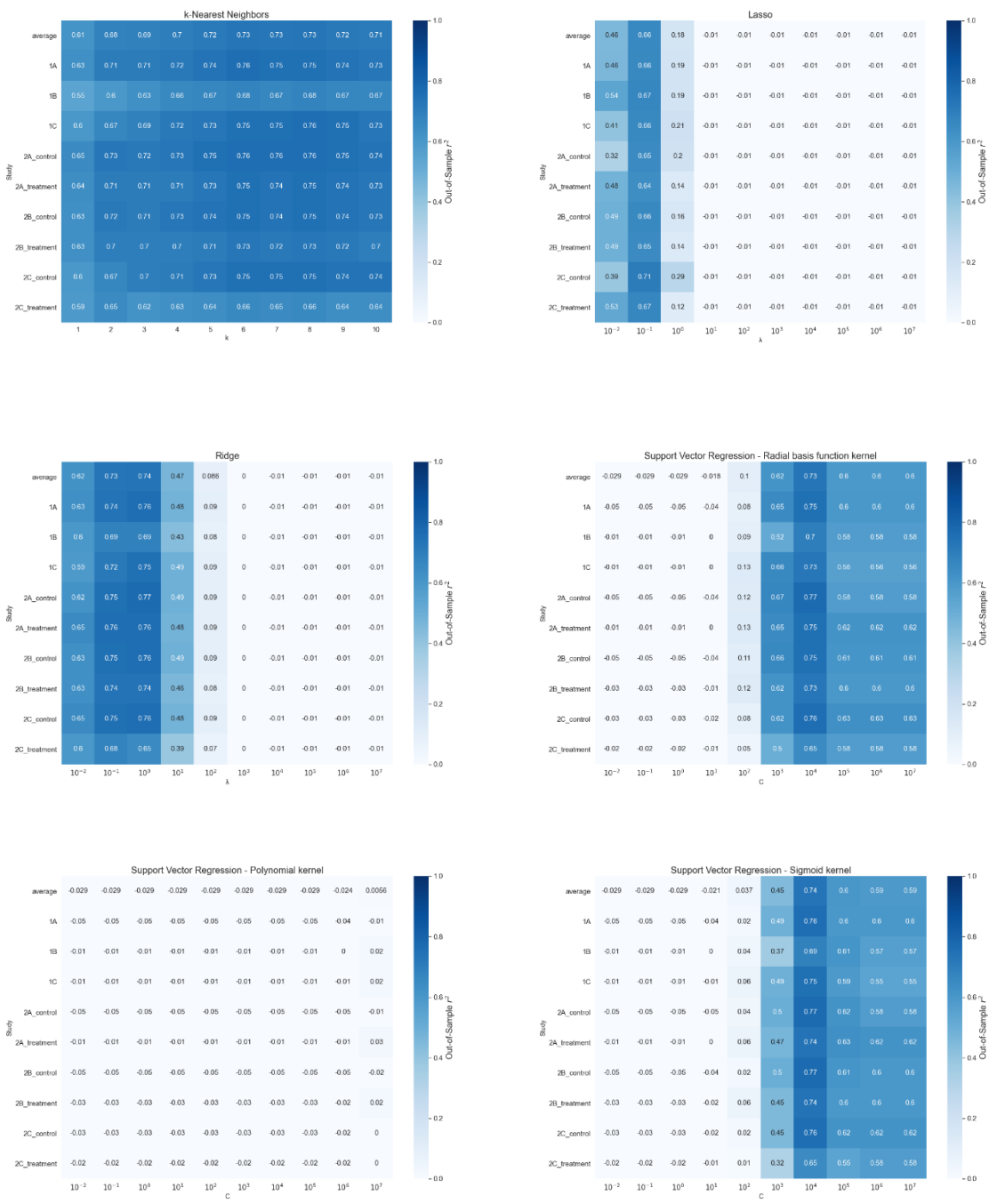
$$\sum_i (y_i - \beta_0 - \sum_j x_{ij} \beta_j)^2 + \lambda \sum_j \beta_j^2 \quad (1)$$

$$\sum_i (y_i - \beta_0 - \sum_j x_{ij} \beta_j)^2 + \lambda \sum_j |\beta_j| \quad (2)$$

Support vector regression uses a “kernel trick” to learn a nonlinear mapping from semantic vector representations to healthiness ratings with the penalization parameter, c , which works similarly to λ in the lasso and ridge techniques. We considered three common kernel functions – radial basis function kernel (SVR-RBF), polynomial kernel (SVR-Polynomial), and sigmoidal kernel (SVR-Sigmoid). K-nearest neighbor regression predicts the healthiness rating of a food item by computing the average rating of the k nearest food items in the semantic vector space. The optimal penalization parameters λ (in lasso and ridge) and c (in SVRs) and the optimal number of neighbors k are chosen through leave-one-out cross-validation. All analyses were implemented in the Python scikit-Learn machine learning library (Pedregosa et al., 2011). For simplicity, keeping other tuning parameters in this library as default, we tested penalization parameters, λ (in lasso and ridge) and c (in SVRs), in the following set: $\{10^{-2}, 10^{-1}, 10^0, 10^1, 10^2, 10^3, 10^4, 10^5, 10^6, 10^7\}$, and number of neighbors, k , in the following set: $\{1, 2, 3, \dots, 10\}$. Figure S2 shows out-of-sample r^2 of different regression techniques with different tuning parameters for every study.

Figure S2

Out of Sample Accuracy of Secondary Vector Representation Models

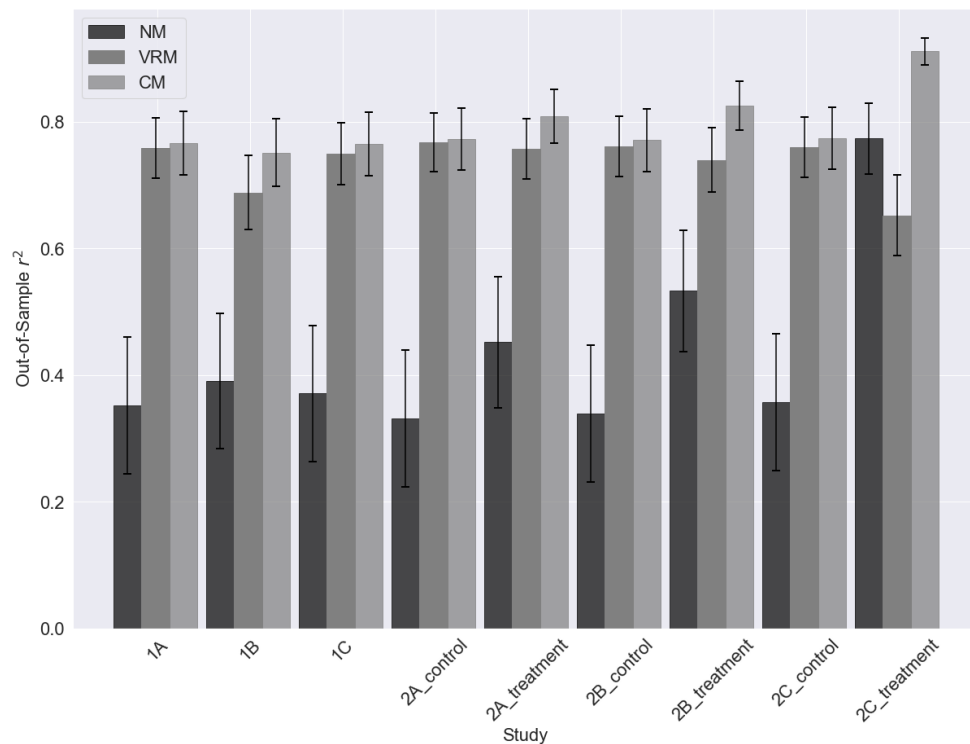


Note. Out-of-sample r^2 of secondary Vector Representation Models with different penalization parameters.

6. Test for Model Comparison

Figure S3 shows the out-of-sample r^2 with the 95% CI for the Nutrient Model, the Vector Representation Model, and the Combined Model trained on average healthiness ratings. Table S6 reports the Pearson correlation between the observed average healthiness ratings and the leave-one-out predictions, the out-of-sample r^2 , and the mean squared errors of the Nutrient Model, Vector Representation Model, and the Combined Model. We can see clearly that both the Vector Representation Model and the Combined Model significantly outperformed the Nutrient Model except in the treatment condition of Study 2C where participants were given the most amount of nutrient information. We also ran separate paired sample t-tests to compare the squared errors from different models for each study. Across all studies, the mean squared errors from the Vector Representation Model and the Combined Model were significantly lower than those from the Nutrient Model (all $p < 0.01$). These results hold for the extended versions of the Nutrient Model and the Combined Model (detailed statistics are available upon request), suggesting that both the Vector Representation Model and the Combined Model outperformed the Nutrient Model.

Overall, the mean squared errors from the Vector Representation Model were not significantly different from those from the Combined Model (and its extended version, detailed statistics are available upon request), suggesting that adding nutrient information (on either macro- or micro-level) to the word vectors did not significantly improve the predictive accuracy of the Vector Representation Model.

Figure S3*Group Level Modeling: Model Comparisons*

Note. Out-of-sample r^2 of the Nutrient Model (NM), Vector Representation Model (VRM), and Combined Model (CM). Error bars represent the 95% CIs.

Table S6

Pearson correlation r , out-of-sample r^2 and mean squared errors (MSE) for the Nutrient Model, Vector Representation Model, and Combined Model

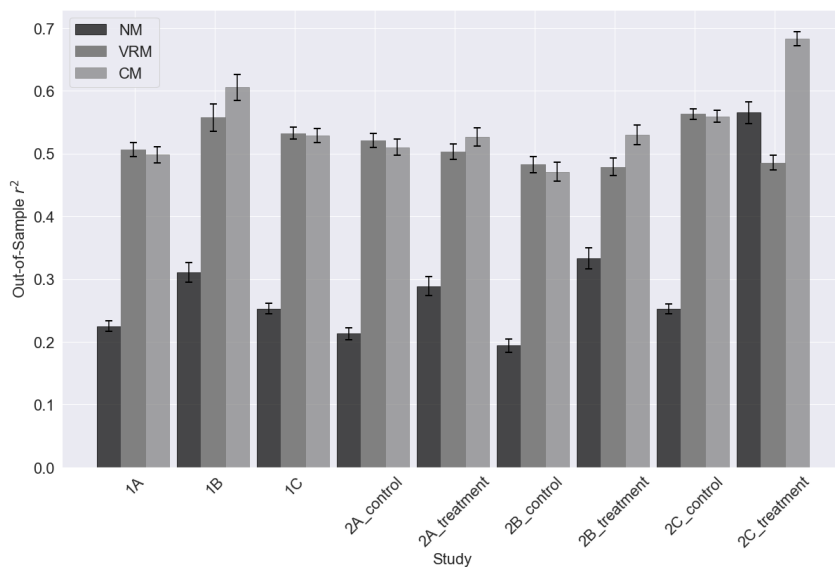
Study	Nutrient Model			Vector Representation Model			Combined Model		
	r	r^2	MSE	r	r^2	MSE	r	r^2	MSE
1A	0.60	0.35	1295.48	0.88	0.76	482.76	0.88	0.77	467.30
1B	0.63	0.39	1471.25	0.84	0.69	753.24	0.87	0.75	600.55
1C	0.61	0.37	1207.45	0.88	0.75	480.31	0.88	0.76	451.33
2A_control	0.58	0.33	1353.49	0.89	0.77	471.76	0.88	0.77	460.46
2A_treatment	0.67	0.45	1002.48	0.88	0.76	443.73	0.90	0.81	349.52
2B_control	0.59	0.34	1183.60	0.88	0.76	427.76	0.88	0.77	410.59
2B_treatment	0.73	0.53	828.56	0.87	0.74	461.70	0.91	0.83	309.75
2C_control	0.60	0.36	1663.91	0.88	0.76	622.98	0.88	0.77	586.05
2C_treatment	0.88	0.77	520.13	0.82	0.65	800.98	0.95	0.91	204.39

7. Individual-level Modeling

In the main text, we evaluated a Nutrient Model, Vector Representation Model, and Combined Model on group-level data, meaning that we averaged food healthiness ratings across all participants within each condition. However, since averaging ratings removes noise and variability, it is necessary to check if our results still hold on the individual level. To model individuals with the three types of models, we performed leave-one-out cross-validation on each individual's set of judgments. This yielded an out-of-sample prediction for each rating a participant gave. We then calculated the coefficient of determination (r^2) between these predictions and an individual's actual ratings and averaged these calculations across participants within studies. Figure S4 visualizes the out-of-sample r^2 for the three types of models for each study. The results on the group level are largely replicated on the individual level. First, the Vector Representation Model performs better than the Nutrient Model across all studies and conditions, except in the treatment condition of Study 2C. Second, the performance of the Nutrient Model increases as more nutritional information is provided in the treatment conditions of Studies 2A-2C. Third, the performances of the Vector Representation Model and Combined Model are indistinguishable in the general public sample of Study 1A and the control condition of Studies 2A-2C. Lastly, the Combined Model performs better than the Vector Representation Model in the expert sample of Study 1B and the treatment conditions of Studies 2A-2C.

Figure S4

Individual Level Modeling: Model Comparisons



Note. Average individual level out-of-sample r^2 of the Nutrient Model (NM), Vector Representation Model (VRM), and Combined Model (CM). Error bars represent standard errors.

8. Food Category Modeling

As a follow-up analysis, we applied our modeling to distinct categories of foods (e.g. baked products, beef products, dairy, and egg products). Overall, our results closely resemble our key findings reported in the main text, as can be seen in Table S7. The Vector Representation Model performs consistently high, with r^2 ranging from 0.76 (95% CI [0.72, 0.81]) to 0.83 (95% CI [0.79, 0.86]). By comparison, the predictive accuracy of the Nutrient Model was much lower but improved with the inclusion of more front-of-pack information across the studies (r^2 ranging from 0.27, 95% CI [0.16, 0.37] to 0.79, 95% CI [0.74, 0.84]). The Nutrient Model performed better than the Vector Representation Model when participants were shown traffic light labeling (Study 2C). However, the Combined Model achieves an even higher predictive ability than any individual model, with r^2 reaching 0.95 (95% CI [0.94, 0.96]) for the most informative (traffic light) labeling system.

We can also note the categories of foods that each of the models performs the best and worst in. There were no substantial differences between the predicted ratings and actual ratings for the Vector Representation Model in all the conditions except when judgments were made in the presence of traffic light labeling. In this condition, the biggest discrepancy between the model's judgment predictions and actual participant judgments was for the "Fats and Oils" category. Irrespective, both model predictions and observed ratings were always less than zero, meaning that the model accurately predicted that participants would perceive foods in this category as unhealthy. The Nutrient Model had the biggest discrepancy for the food categories of "Beverages", "Fats and Oils", and "Nuts and Seed Products", which was apparent in multiple conditions. As an example, for "Beverages", the Nutrient Model predicted healthier ratings than the actual ratings provided by participants in every case. This gap was still present when participants made judgments using traffic light labeling, despite the high overall predictive ability of the Nutrient Model in this condition. A possible explanation is that beverages like "cola" and "beer" do have a low negative nutrient content (e.g. fat, saturates, salt), hence why the model predicted a relatively high healthiness (above a rating of zero). However, in every condition except traffic light labeling, participants rated the "Beverages" category as unhealthy implying a distinct disparity between nutrient content and judgment formation. On the other hand, our Vector Representation Model always had a very high predictive ability for the category of "Beverages" demonstrating, particularly in this instance, that associations capture important attributes underlying healthiness judgments.

Table S7

Out-of-sample r^2 comparisons between the Vector Representation Model, Nutrient Model and Combined Model for food category modeling, numbers in the brackets represent 95 % CIs

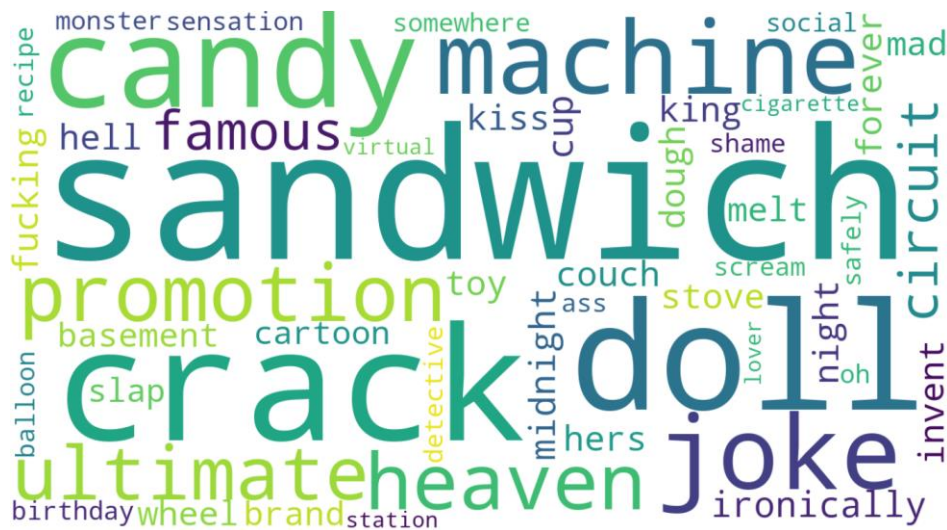
Study	Vector Representation Model	Nutrient Model	Combined Model
1A	0.83 [0.79, 0.86]	0.31 [0.20, 0.41]	0.84 [0.81, 0.88]
1B	0.78 [0.74, 0.82]	0.34 [0.25, 0.43]	0.82 [0.79, 0.86]
1C	0.80 [0.76, 0.84]	0.28 [0.18, 0.39]	0.81 [0.77, 0.85]
2A_control	0.80 [0.76, 0.84]	0.28 [0.17, 0.38]	0.82 [0.78, 0.86]
2A_treatment	0.79 [0.75, 0.84]	0.44 [0.35, 0.53]	0.86 [0.83, 0.89]
2B_control	0.82 [0.79, 0.86]	0.27 [0.16, 0.37]	0.83 [0.79, 0.86]
2B_treatment	0.76 [0.72, 0.81]	0.55 [0.46, 0.64]	0.88 [0.86, 0.91]
2C_control	0.82 [0.78, 0.86]	0.32 [0.21, 0.43]	0.84 [0.80, 0.87]
2C_treatment	0.68 [0.62, 0.74]	0.79 [0.74, 0.84]	0.95 [0.94, 0.96]

9. Word Cloud for Unhealthy Associations

It is possible to identify regions in vector space associated with unhealthiness in the same manner that we uncovered words associated with healthiness in the main manuscript (our Figure 4). As seen in Figure S5, the words with the lowest healthiness ratings (using a model trained on Study 1A participant ratings), have connotations that are more ambiguous. Despite this, there are recognizable terms related to less healthy categories of foods (e.g. “sandwich”, “melt”, “candy”, “alcohol”), and also words that seem to correspond to aspects of the context in which foods can be consumed (e.g. “midnight”, “couch”, “birthday”, “social”). While less intuitive, these associations that allude to the emotional and social benefits of unhealthy foods may explain why consumers choose them over options that are more nutritious.

Figure S5

Unhealthiest “Other” Words based on the Vector Representation Model



10. Test of Model Generalizability

In order to establish whether our Vector Representation Model can provide true out-of-sample predictions about healthiness ratings, we collected participant ratings for an additional (not original 172) 60 foods (e.g. blackberries, custard, and spinach). To obtain this set of new stimuli, we generated model predictions (trained on Study 1A aggregate ratings) for all foods from the Food and Nutrient Database for Dietary Studies (U.S. Department of Agriculture, 2019). We then removed foods included in the original stimuli and uncommon or ambiguous foods. Following this, we selected 20 foods from the lowest quartile, 20 from around the median and, 20 from the highest quartile of predicted healthiness ratings, resulting in our final list of 60 foods. One hundred and one participants were recruited from Prolific Academic, with four participants removed using the same exclusion as the previous studies. There were 99 participants in our final participant sample (aged 19-61 years, $M_{\text{age}} = 26.71$ years, $SD = 8.98$, 39.2% females, and 87.6% had no dietary restrictions). The design of this study was identical to Study 1A. The only difference was that participants rated the healthiness of food names that had not previously been used as stimuli. A paired-sample t-test was run to compare model predictions (trained on Study 1A data) with human ratings on both the aggregate and individual levels. On the aggregate level, the model predictions were not significantly different from the average participant ratings ($t(59) = -0.34$, $p = 0.735$). On the individual level, ratings from 53 out of 99 (53.54%) participants were not significantly different from the model's predictions. Based on these

results, we conclude that the model can be used to predict subjective healthiness judgments for new foods, even in the absence of any nutritional information in the model specification.

11. References

- Mikolov, T., Grave, E., Bojanowski, P., Puhersch, C., & Joulin, A. (2018). Advances in pre-training distributed word representations. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *NeurIPS*, 3111-3119.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, 12(Oct), 2825-2830.
- Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532-1543.
- U.S. Department of Agriculture, Agricultural Research Service. 2019. USDA Food and Nutrient Database for Dietary Studies 2017-2018. Food Surveys Research Group Home Page, <http://www.ars.usda.gov/nea/bhnrc/fsrg>