Supplementary Note 1: Pre-registered analyses and deviations from pre-registered analysis plan

The pre-registered analysis plan for this study is available at https://osf.io/7krnt/.

Here, we list the four pre-registered primary research questions and the sections of this document that report how we answered them (including both pre-registered and complementary non-pre-registered analysis methods):

Research question 1. Will the intervention change adolescents' construals of the values-alignment of healthy eating and the social status appeal of healthy eating?

- This analysis was conducted as pre-registered and is reported in the Supplementary Results 1 section of this document and in Supplementary Figure 1.

Research question 2. Will the intervention cause adolescents to select fewer junk foods (with fewer grams of carbohydrates) on a surreptitious "snack pack" form?

- This analysis was conducted as pre-registered and appears in the Supplementary Results 1 section of this document of this document.

Research question 3. Will the intervention change explicit and implicit perceptions of food marketing materials?

- Explicit attitudes: This analysis was conducted as pre-registered and appears in the Supplementary Results 1 section of this document and in Supplementary Figure 1.
- Implicit attitudes:
    - The pre-registered analysis (independent t-tests for each measurement occasion) was conducted as planned and appears in Supplementary Results 1. Upon further consideration, however, we deemed this analysis to be less appropriate for reasons explained in that section.
    - The complementary non-pre-registered analysis (a repeated measures analysis, which we deemed to be more appropriate given the nature of the data) also appears in Supplementary Results 1 and in Supplementary Table 1.

Research question 4. Will the intervention change daily snack and drink purchases in the cafeteria over time?

- This analysis was conducted as described in the pre-registration. It appears in Supplementary Results 1 and in Supplementary Table 2.
- An exploratory analysis discovered strongly significant moderation by gender, and this is reported in Supplementary Results 1 and in Supplementary Table 3.
- Because the moderation by gender was not predicted or specified in advance, a permutation test, reported in Supplementary Results 2, was used to probe the robustness of this result.

Several follow-up tests of the robustness of the results were not pre-registered and appear below:

- An examination of the effectiveness of random assignment appears in Supplementary Results 2.
- An examination of the potential effects of attrition and differential attrition on our inferences also appears in Supplementary Results 2 and in Supplementary Table 4.

- An examination of whether the results for the primary research questions are robust to the inclusion of interactions between the experimental condition variable and demographic characteristics that could have shown differential attrition is reported in Supplementary Results 2 and in Supplementary Table 5.

Two follow-up analyses were listed as secondary and exploratory in the pre-registration document but deemed inappropriate for the present paper upon further consideration:

- We do not report a test of whether explicit and implicit attitudes toward food marketing mediated the intervention's effect on cafeteria choices because we realized, after completing the pre-registration, that the correct test of that theoretical idea (i.e., that changing adolescents' affective associations with food marketing mediates the effect of the intervention on cafeteria choices) actually is a moderated mediation test. Individuals' negative affective associations with food marketing should "boost" the treatment effect to the extent that individuals are exposed to such marketing. Because data on participants' exposure to food marketing were not available in this study (and because such exposure might even be endogenous to the treatment effect), we suggest that this question is best addressed in future research on the mechanisms that sustain the effects of values-alignment messaging.
- We do not report exploratory moderation analyses by testosterone level and other indicators of pubertal development. It was not clear whether we would have those data available (due to costs of hormone assays), and so we indicated in the pre-registration document for this study that we would pre-register any hypotheses about those tests separately. That pre-registration has been completed. It can be found at https://osf.io/6j85b/ and will be the subject of its own report.

Supplementary Note 2: Sample, statistical power, and checks on internal validity

Characteristics of the participating sample of students. Here are the demographic characteristics of the participating students:

- *Age*: 13 years old = 44% (n = 158), 14 years old = 52% (n = 189), 15 years old = 4% (n = 15)
- *Gender*: Male = 49% (n = 165), Female = 51% (n = 175)
- *Race*: White, non-Hispanic = 51% (n = 183), Other racial/ethnic group membership = 49% (n = 179)
- *Economic disadvantage status*: Reduced-price lunch = 31% (n = 104), Free lunch = 7% (n = 24)
- *Overweight (BMI>85th %ile)*: 26% (n = 66)
- *Obese (BMI>95th %ile)*: 11% (n = 28)

No student who was present on the day of the intervention declined to participate in the study, and so these values match the characteristics of the school as a whole.

Statistical power and determinants of sample size. As we indicated in our pre-registration document, our sample size was determined with a "stopping rule": we collected data from all students at our partner school who were enrolled in the 8th grade and in school on the first day of the study (*N* = 362). Assessing the statistical power this sample size afforded us for our two primary outcomes (implicit affective associations and cafeteria purchases) is a challenge because we do not have a strong basis on which to estimate what a likely effect size might be. Relying on our observed effect sizes would yield power analyses that are completely redundant with the *P*-values we already report[1] and, since we are not aware of any previously documented instances of effects on either of our outcomes lasting anywhere near as long after the conclusion of the treatment as we observe here, we lack good external sources to guide such an estimate. We are aware, however, of one instance of an experimental effect on implicit attitudes that was observed at $p = 0.024$, $d = |0.34|$, at two days post-intervention (see Ref. [2], p. 1006). This effect did not meet that paper's threshold for statistical significance because it was one of a large number of comparisons tested, however it is the strongest change in implicit attitudes we are aware of that has been documented beyond the same day as an experimental treatment (and the only such instance to reach any conventional standard of statistical significance) so we use it to guide our estimate. Our sample afforded us 86% power to detect an AMP effect as large or larger than that.

We could find no comparable exemplar to guide a power analysis for this first study of cafeteria purchases, because (a) the average effect on boy's dietary preferences for past interventions is not different from zero, (b) no past study we could find for any age group or gender has reported effect sizes coming from data with the structure of the data we collected (i.e., individual-level daily consumption data), and (c) standardized effect sizes and power analysis for cafeteria purchase are dependent on parameters that are likely to be idiosyncratic to a given school and hard to predict in advance, such as the distribution of purchasing choices per person per day and the within-person correlations of purchasing decisions over time. We did design our approach to statistical modelling, however, to take advantage of the repeated measures element of those data and maximize statistical power. Moreover, one primary reason to be interested in statistical power once a study is known to have detected statistically significant effects is to assess the likelihood of a Type 1 error (under the assumption that under-powered studies showing significant effects might be over-interpreting noise in the data. To minimize this likelihood, we fixed the sample size before we had access to the data (as described above) and we pre-registered our analysis plan.

The only substantive result we report and interpret in the paper that was not in the pre-registration plan was our discovery of moderation of the condition effect on cafeteria purchases by gender (not including robustness checks). Therefore, we report an assessment of the robustness of this effect (a permutation test) in the Supplementary Results 2 section. It further supports our assertion in the main text that the moderation patterns for gender are unlikely to be caused by noise in the data.

Effectiveness of random assignment. As expected, participants randomly assigned to the exposé intervention did not differ significantly from the participants randomly assigned to the control intervention in terms of characteristics we measured:

- *Age*: $\chi^2(2) = 0.73$, $P = 0.693$, $N = 362$
- *Gender*: $\chi^2(1) = 0.05$, $P = 0.823$, $N = 340$
- *Race*: $\chi^2(1) = 1.88$, $P = 0.171$, $N = 362$
- *Economic disadvantage status*: $\chi^2(2) = 0.16$, $P = 0.924$, $N = 340$
- *Teacher*: $\chi^2(2) = 0.12$, $P = 0.941$, $N = 362$
- *Overweight BMI*: $\chi^2(1) = 0.00$, $P > 0.999$, $N = 250$

- *Obese BMI*: $\chi^2(1) = 0.00$, $P > 0.999$, $N = 250$

Fidelity to protocol, by intervention condition. As pre-registered, all analyses of the effects of the exposé intervention were conducted with the intent-to-treat sample, which means that participants' data were included regardless of whether they successfully completed the intervention exercises. This is essential for maintaining valid causal inference.

The mechanism for an intervention effect, however, can be obscured if different conditions evoke different rates of engagement with the experimental materials. Therefore, on an exploratory basis, we examined whether the two experimental conditions might have differed in their fidelity to the intervention.

The analyses detailed below show that the participants randomly assigned to the exposé intervention did not differ significantly from the participants randomly assigned to the control in terms of the percentage of participants who successfully completed the materials:

- *Completion of the first essay assignment in Session 1*: $\chi^2(1) = 1.08$, $P = 0.299$, $N = 362$
- *Completion of the second essay assignment in Session 1*: $\chi^2(1) = 0.09$, $P = 0.766$, $N = 362$
- *Beginning Session 2*: $\chi^2(1) = 0.01$, $P = 0.904$, $N = 362$
- *Completion of Session 2 "booster" video*: $\chi^2(1) = 0.33$, $P = 0.567$, $N = 362$

Supplementary Results 1: Detailed analysis methods for substantive research findings

Research question 1: Will the exposé intervention change adolescents' construals of the values-alignment of healthy eating and the social status appeal of healthy eating? (pre-registered analyses)

The pre-registered analysis plan stated that we would conduct simple t-tests to compare the group means and answer Research Question 1. The pre-registered analyses showed the predicted, significant difference between the conditions both for values-aligned construal (i.e. autonomous and social-justice-oriented) and social status appeal:

- *Values-aligned construal*: Exposé $M = 3.52$, $SD = 0.88$, $N = 173$; Control $M = 2.55$, $SD = 0.86$, $N = 167$, $t(338) = 10.23$, $P < 0.001$, $SMD = 1.12$, 95% CI_{SMD}: [0.90, 1.33]. The results are substantively identical when considering the two aspects of the values-aligned construal (autonomy and social justice) separately.
- *Social status appeal*: Exposé $M = 3.49$, $SD = 0.87$, $N = 173$; Control $M = 3.04$, $SD = 0.93$, $N = 168$, $t(339) = 4.66$, $P < 0.001$, $SMD = 0.49$, 95% CI_{SMD}: [0.28, 0.70].
- *Mediation analysis*: A mediation analysis, following current recommendations for best practices[3], documented a significant indirect effect of the exposé intervention on social status appeal via values-aligned construal, $b = 0.64$, 95% CI_b: [0.53, 0.76].

Research question 2. Will the exposé intervention cause adolescents to select fewer junk foods (with fewer grams of carbohydrates) when ordering a "snack pack" treat provided by the school principal 1-week post-intervention? (pre-registered analyses)

In the present study, we administered the "snack pack" order form one week after the intervention while, in our past research[4], we did so the day after the intervention.

The pre-registered analysis plan stated that we would conduct simple t-tests to compare the group means and test Research Question 2. Contrary to predictions, in the present study, when the “snack pack” form was administered one week post-intervention, we found no significant condition differences on “snack pack” order form choices, either in terms of *number* of junk choices or *carbohydrate content*. The analysis plan also stated that we could conduct ordered probit analyses of the snack form, but the results were the same and so we only present the t-test estimates.

- *Number of junk choices*: Exposé M = 2.38, SD = 0.77, N = 156; Control M = 2.32, SD = 0.78, N = 139, $t(293) = -0.75$, P = 0.45, SMD = 0.08, CI_{SMD}: [-0.14, 0.32].
- *Total carbohydrate content of junk choices*: Exposé M = 58.10, SD = 20.77, N = 156; Control M = 56.17, SD = 20.56, N = 139, $t(293) = -0.80$, P = 0.43 SMD = 0.09, CI_{SMD}:[-3.65, 8.58].

In sum, there were no significant intervention effects on one of the pre-registered outcomes: the “snack pack” form.

The “snack pack” form was the primary outcome in our past research[4]. The present research is not a replication of that study, because the snack form was administered 1 week, not 1 day, post-test.

Speculating, the longer delay in measurement could explain the null finding for two reasons. First, since the form only represents one choice, it is likely to have substantially more measurement error than the other measure of food choices (daily cafeteria purchases). If the exposé message was less focal in memory after a week, that could interact with greater measurement error to mask treatment effects. Second, the snack pack was framed by the school principal as a celebration for good behaviour. Participants may therefore have thought of the snack pack as a special occasion rather than as a reflection of their daily habits. Although that was likely also true in the initial evaluation study, the temporal proximity of the snack pack measure to the intervention in that study might have allowed an effect to manifest anyway. For example, in the present study, having established a weeklong pattern of healthier choices in their daily cafeteria purchases might have made participants more likely to indulge in their snack pack choices without concern that this indulgence was reflective of their overall inclination.

Research question 3. Will the exposé intervention change explicit and implicit perceptions of food marketing materials?

Explicit attitudes: Anger in response to viewing junk food advertisements and desire to consume junk food (pre-registered analyses). As stated in the pre-registered analysis plan, we conducted simple t-tests to answer Research Question 3.

- *Anger in response to junk food advertising*: Exposé M = 2.28, SD = 1.16, N = 156; Control M = 1.34, SD = 0.85, N = 149, $t(303) = 8.00$, $P < 0.001$, SMD = 1.09, CI_{SMD}: [0.82, 1.37].
- *Desire to consume junk food products*: Exposé M = 2.45, SD = 1.29, N = 157; Control M = 2.87, SD = 1.33, N = 149, $t(304) = -2.85$, P = 0.005, SMD = 0.32, CI_{SMD}: [0.10, 0.54].

Implicit attitudes: Responses to junk and healthy food in the Affect Misattribution Procedure (AMP) (overview). Participants completed measures of their implicit attitudes toward junk food advertising and healthy food images, measured using the Affect Misattribution Procedure (AMP), immediately following the intervention, at two-week (March) follow-up, and at three-month (May) follow-up. The school's Wi-Fi connection failed during the immediate post-task follow-up (February) for most participants. As a result, data from the February AMP are not definitive on their own. But those results are consistent with the AMP data from the other two time points, as shown below. Our conclusions are substantively unchanged if the February AMP data are excluded. (see additional analyses below, Supplementary Figure 2)

We report two methods of analysis of the AMP. First, we report a repeated-measures analysis in which AMP responses to healthy and junk food images are nested within individuals. This is preferable if there are no differences across measurement occasions in the effects of the treatment. Second, we report simple t-tests comparing exposé and control condition participants in terms of their composite AMP responses (junk food 'pleasant' ratings minus healthy food 'pleasant' ratings) for each time point. The latter was our pre-registered analysis. However, upon reconsideration, we believe the repeated-measures analysis is more appropriate because, by combining all data, it allows for superior statistical power. Confirming this decision, analyses find no interactions with stimulus or time (see Supplementary Table 1), so it is more appropriate to combine all time points.

Finally, we note that, on an exploratory basis, we also assessed implicit attitudes toward a third kind of stimulus, not analysed here: pictures of unhealthy foods sold in the school cafeteria. We did not pre-register the cafeteria-specific stimuli because they were not the primary interest.

<u>AMP implicit attitudes analysis #1: Linear mixed-effects models (non-pre-registered analysis).</u> In the repeated measures analysis, we reverse-coded the proportion saying that healthy options were "pleasant," so that it was on the same metric as the junk food options (Model 1 in Supplementary Table 1). As a robustness test, we next conducted analyses of the healthy and junk food stimuli independently (Models 2 and 3 in Supplementary Table 1). The repeated measures analysis of AMP responses was estimated via a mixed effects model with responses nested within participants. It took the following general form (represented using the lmer function in R[5]):

```
lmer(amp_pleasant_responses~Condition*Type*Time+(1|RAND), data=amp_long)
```

Analyses showed an overall effect of exposé intervention condition assignment on AMP responses when combining unhealthy stimuli and healthy stimuli (reverse-scored), $b = -0.058$, $SE = 0.016$, $t(375.651) = -3.649$, $P < 0.001$, $SMD = 0.23$, 95% CI_{SMD}: [0.11, 0.36]. Furthermore, this condition difference was significant for both healthy stimuli (coded such that higher numbers represent a higher proportion of "pleasant" responses), $b = 0.047$, $SE = 0.021$, $t(347.112) = 2.283$, $P = 0.023$, $SMD = 0.20$, 95% CI_{SMD}: [0.02, 0.38], and junk food (coded such that higher numbers meant more "pleasant" responses), $b = -0.068$, $SE = 0.026$, $t(348.729) = -2.645$, $P = 0.009$, $SMD = 0.26$, 95% CI_{SMD}: [0.06, 0.45]. These reported coefficients are from a model that centres the time and type variables at the sample mean. Supplementary Table 1 reports coefficients from a model in which the type and time variables are dummy coded.

In addition to the categories of AMP stimuli described in our pre-registered analysis plan, we included a set of images, for exploratory purposes, of the specific unhealthy snacks and drinks available at that school cafeteria. Analyses of AMP responses to those images yielded substantively similar results to analyses of the junk food images: *b* = -0.053, *SE* = 0.025, *t*(345.985) = -2.137, *P* = 0.033, *SMD* = 0.21, CI$_{SMD}$: [0.01, 0.41].

As noted, the intervention effect was not different across measurement occasions (see non-significant interactions of condition assignment with time in Supplementary Table 1). This suggests that the exposé intervention changed AMP responses equally effectively for each class of stimulus and that there was no significant decay in intervention effects over the 3-month follow-up period. Also see Supplementary Figure 2, which depicts response distributions by stimulus and measurement occasion; these show a consistent pattern of treatment effect. (Our primary description of the AMP results, in the paragraph immediately above, describes the main effect of the exposé intervention condition without higher-order interactions).

<u>AMP implicit attitudes analysis #2: Separate t-tests (pre-registered analysis).</u>

- *February (immediate) AMP responses*: Exposé *M* = -0.43, *SD* = 0.45, *N* = 20; Control *M* = -0.24, *SD* = 0.43, *N* = 37, *t*(55) = 1.61, *P* = 0.113, *SMD* = 0.44, 95% CI$_{SMD}$: [-0.13, 1.00] (recall that this result, in isolation, should be treated with caution due to missing data).
- *March (two-weeks post-intervention) AMP responses*: $M_{Exposé}$ = -0.19, *SD* = 0.34, *N* = 155; $M_{Control}$ = -0.06, *SD* = 0.29, *N* = 152, *t*(305) = 3.56, *SMD* = 0.37, 95% CI$_{SMD}$: [0.15, 0.60].
- *May (three-months post-intervention) AMP responses*: Exposé *M* = -0.18, *SD* = 0.33, *N* = 169; Control *M* = -0.09, *SD* = 0.29, *N* = 161, *t*(328) = 2.58, *P* = 0.011, *SMD* = 0.27, 95% CI$_{SMD}$: [0.05, 0.48].

Research question 4. Will the exposé intervention change daily snack and drink purchases in the cafeteria?

To answer Research Question 4, we examined the effect of the exposé intervention on changes in participants' purchases in the cafeteria for the remaining three months of the school year after the intervention. Prior to researchers receiving the cafeteria purchase data from the school, professional research staff, who were blind to participants' condition assignment, categorized all choices into two categories: healthy snacks and junk food snacks or drinks. The **pre-registered, confirmatory analyses** focused on the predicted increase in healthy snacks and the predicted decrease in junk food snacks and drinks in the period after the exposé intervention.

The statistical model for estimating treatment effects on changes in daily purchases took the following general form (represented using the lmer function for R[5]):

```
lmer.test(lmer(junkpurchases~prepost*treatment+
     date+sex+age+pre.treatment.purchases+
     (1|person), data=data))
```

To understand *changes* in purchasing behaviour, we estimated a Time (0 = Pre-intervention, 1 = Post- intervention) x Intervention condition (1 = Exposé, 0 = Control) interaction in the linear mixed effects model.

To ensure that chance differences in observed baseline characteristics or behaviours could not explain the intervention effect, analyses controlled for participant sex, age, and total number of pre-intervention purchases. Analyses that did not control for these variables yielded the same substantive conclusions.

The primary model was a linear probability model (1 = Purchase that day, 0 = No purchase that day). When conducting the analyses using a mixed effects logistic model in Stata, the conclusions were identical (but the model took longer to estimate).

Analysis showed intervention effects (i.e., Intervention condition x Time interactions) that were significant at the $P < 0.05$ level and in the theoretically expected direction of showing reductions in unhealthy choices and increases in healthy choices (see Supplementary Table 2).

Gender interactions in exposé intervention effects on lunch purchases (exploratory analyses). Exploratory analyses, using linear mixed effects regression models, revealed a significant gender interaction, showing statistically significant differences between males and females in terms of changes from pre-intervention to post-intervention (see Supplementary Table 3, Row 8). Odd-numbered models (1, 3) in Supplementary Table 3 code the gender variable so that males are the reference group, and therefore in those columns, Row 5 depicts the Intervention condition x Time interactions among males. These models show that males' purchasing behaviours changed in the direction of healthier purchases from pre- to post-intervention. That is, the exposé intervention reduced boys' junk food purchases and increased their healthy purchases.

Even-numbered models (2, 4) in Supplementary Table 3 code the gender variable so that females are the reference group and depict the Intervention condition x Time interactions among females. Row 5 in Supplementary Table 3 shows that females seemed to show slightly stronger reductions in junk food in the control relative to the exposé condition, which was not expected.

Note that gender analyses are only possible with the sub-sample of participants for whom gender data were available either from self-reports or from the school district. (Gender data were missing for 22 participants.)

Analysis of time trends reported in Supplementary Table 3 reveals that there was a general trend toward unhealthy choices as the school year progressed (i.e. more junk, less healthy food) for all participants except for boys who received the exposé intervention.

Calculating relative change in cafeteria purchases for boys. Among boys, the estimated likelihood of buying unhealthy items in the pre-intervention period was 0.143 unhealthy snacks or drinks per person per day (therefore, the reduction of 0.044 = 31%). The likelihood of buying a healthy snack or drink (e.g., water, fruit) for boys in the pre-intervention period was very low: 0.005 healthy snacks per person per day. For that reason, the daily absolute increase in healthy purchases of 0.008 healthy purchases per person per day corresponds to a very large percentage increase: 160%. The reference to a 35% overall improvement in the health profile of boys' purchases in the main text is based on summing the absolute value of the reduction in junk food purchases (0.44) and the increase in healthy purchases (0.008) and computing that sum as a percentage of the sum of baseline junk food (0.143) and healthy (0.005) purchases. (See Supplementary Table 3.)

Frequency with which students made purchases in the cafeteria. Ninety three percent of students made a purchase using their ID cards on at least one day, 86% made a purchase on more than 5% of days, and most students (approximately 65%) made a purchase on a majority of days (see Supplementary Figure 2 for the complete distribution).

Supplementary Results 2: Sensitivity analysis and tests of robustness

Sensitivity analysis: Could the effects of the exposé intervention be explained by differential attrition?

Missing data. In accordance with recommendations from the U.S. Department of Education's Institute for Educational Sciences (IES) technical report on missing data in the evaluation of school-based interventions, we (a) did not impute missing data for outcome variables and (b) used the missing data dummy variable approach to handle missing data for covariates.[6]

Levels of attrition by condition. According to another IES technical report on the evaluation of school-based interventions, small amounts of attrition can bias the estimate of the treatment effect even when the treatment is randomly assigned[7]. Attrition is therefore a potential threat to internal validity. The WWC has set a benchmark of less than 20% attrition overall and no more than a 7-percentage point difference in attrition rates between conditions to avoid concerns about attrition-related bias in results.

As shown below, attrition rates overall and by condition were well within the WWC-recommended levels for evidence that meet the standard of "without reservations," except in two cases: AMP data at immediate post-test (where loss of data occurred due to a problem with the school's Wi-Fi connection, as discussed above) and the "snack pack" order form.

- *Autonomous and social-justice-oriented construal of healthy eating*: Exposé = 6%, Control = 6%, Difference = 0 percentage points.
- *Explicit reactions to food advertising (i.e. anger, desire to consume)*: Exposé = 15%, Control = 16%, Difference = -1 percentage points.
- *"Snack pack" order form (1 week post-intervention)*: Exposé = 15%, Control = 22%, Difference = -7 percentage points.
- *March AMP implicit attitudes (2 weeks post-intervention)*: Exposé = 16%; Control = 15%, Difference = 1 percentage points.
- *May AMP implicit attitudes (3 months post-intervention)*: Exposé = 16%, Control = 15%, Difference = -2 percentage points.

Thus, measures for which the exposé intervention was found to have significant effects are below the attrition threshold that would raise concerns about possible bias. The longest-term outcome (AMP responses at 3 month follow-up) showed less attrition relative to the medium-term outcomes. It is not necessary to conduct an attrition analysis for the lunchroom purchase data because the school provided complete data.

The "snack pack" order form, however, had 22% attrition in the control condition and a 7-percentage point difference in attrition rates between the two conditions, by far the greatest degree of differential attrition of any measure in the study. It is possible this contributed to the failure of that measure to reveal effects consistent with the more naturalistic cafeteria purchase data.

Characteristics of attritors by condition. Next, we explore the potential for attrition to be *differential*, defined as the possibility that the characteristics of attritors in one condition might differ in meaningful ways relative to the other experimental condition. Supplementary Table 4 presents linear probability models predicting a dummy variable indicating missing data. These found a few differences by condition in terms of the characteristics of attritors, but no more than would be expected due to chance.

Robustness test: Effects controlling for interactions of condition with demographic characteristics

Analyses probed the robustness of the study's experimental effects when controlling for interactions between condition and demographic characteristics. These analyses were only possible in the sub-sample of participants who reported demographic information or whose demographic information could be obtained from school records. All demographics were centred, and so the exposé intervention effect estimate is the sample-average treatment effect. The AMP measure reported here is the net positivity score (the pleasantness of junk food minus the pleasantness of healthy food). Analyses revealed all of the main study analyses we examined are robust to controlling for interactions with demographic factors (see coefficients for "Exposé" in Supplementary Table 5).

Robustness test: Difference-in-differences specification of intervention main effect on cafeteria purchases with fixed effects for individual participants

To gauge the robustness of the intervention's main effect on daily cafeteria purchases, we supplemented the pre-registered mixed-effects model with a popular alternative analytical approach. Following the approach implemented by Allcott[8], who also evaluated the effects of a relatively brief intervention on repeated-measure behavior data, we specified a difference-in-differences estimator, which models daily lunch room purchases as a function of intervention condition, an indicator for pre- versus post-treatment time period, and their interaction, controlling for a linear effect of date before and after the treatment, with fixed effects for individuals. The individual-level fixed effects remove any time-invariant individual differences that could predict lunchroom choices, so it is a useful method for removing the effect of any chance variation between individuals in the sample resulting from failure of random assignment. The model was estimated with the the xtreg function in Stata SE. This analysis yielded substantively identical results to those of the pre-registered mixed-effects model: a significant increase in the rate of healthy purchases, $b = 0.002$, $SE = 0.001$, $z = 2.06$, $P = 0.039$, $OR = 1.49$, 95% CI_{OR}: [1.02, 1.97], and a significant decrease in unhealthy purchases, $b = -0.012$, $SE = 0.005$, $z = 2.21$, $P = 0.027$, $OR = 0.92$, 95% CI_{OR}: [0.85, 0,99] in the exposé condition compared with the control condition.

Robustness test: Permutation test assessing the likelihood that the observed data could have produced the gender moderation results by chance alone

As noted, the only result we report and interpret in the paper that was not in the pre-registration plan was our discovery of moderation of the condition effect on cafeteria purchases by gender.

This exploratory moderator analysis was theoretically motivated; cultural norms in the U.S. place much more pressure on teenage girls than on teenage boys to be thin. Moreover, past research consistently finds stronger effects of obesity-related health interventions for girls than for boys[9,10]. The gender moderation is unlikely to be the result of multiple hypothesis testing because we did not test other moderators prior to finding the gender moderation, because the potential universe of pre-intervention moderators was limited to five variables obtained from administrative sources (age, race/ethnicity, gender, socio-economic disadvantage, and BMI), and because the *P*-values associated with the interactions are robust to even the most conservative Bonferroni corrections for multiple hypothesis testing across these five.

Nevertheless, to provide an additional gauge of the likelihood that a moderation result as extreme as we found would occur by chance, we conducted an analysis known as a "permutation test"[11]. This test complements the reported *P*-value, accounting for the possibility that "noise" in the data (such as extreme observations, or complex structures to the data that deviate from the assumptions of the statistical model) could, by chance, have been allocated to one experimental condition or the other or one gender or the other, potentially causing a "false positive" result for the moderation test. We estimated the likelihood of this by generating 1000 simulated samples in which the condition and gender variables were randomly shuffled (so the null hypothesis is known to be true for the condition effect and its interaction with gender) but all the other variables were held identical to the real data.

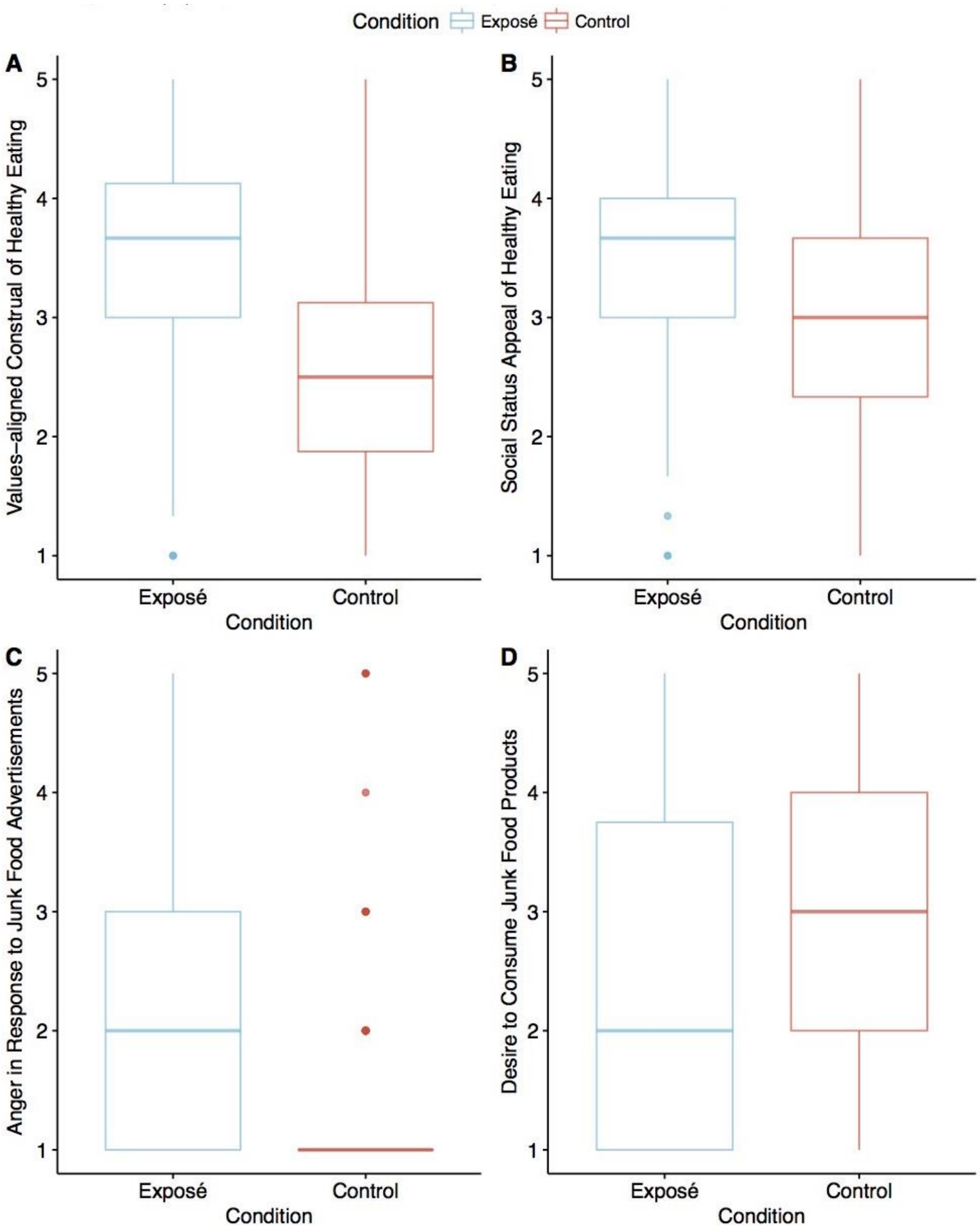
We estimated the size of the interaction of the intervention with gender in each of those 1000 true-null samples and found that only 2.6% yielded an unstandardized interaction effect size between condition and gender as strong as or stronger than the one identified in the real data (in either direction). Only 1.3% of true-null simulated samples yielded an interaction effect as strong or stronger and in the same direction—that is, an interaction favouring boys to the same extent as the real data. Unlike regression or any analysis that relies on distributional assumptions, the permutation test is robust to outliers and other irregularities in the data (like complex failures of random assignment) that can cause spurious significance test results.

Supplementary References

1. Lakens, D. The 20% Statistician: Observed power, and what to do if your editor asks for post-hoc power analyses. *The 20% Statistician* (2014).
2. Lai, C. K. *et al.* Reducing implicit racial preferences: II. Intervention effectiveness across time. *J. Exp. Psychol. Gen.* **145**, 1001–1016 (2016).
3. Imai, K., Keele, L. & Tingley, D. A general approach to causal mediation analysis. *Psychol. Methods* **15**, 309–334 (2010).
4. Bryan, C. J. *et al.* Harnessing adolescent values to motivate healthier eating. *Proc. Natl. Acad. Sci.* **113**, 10830–10835 (2016).
5. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models using lme4. *ArXiv14065823 Stat* (2014).
6. Puma, M., Olsen, R. B., Bell, S. H. & Price, C. What to Do When Data Are Missing in Group Randomized Controlled Trials. 131
7. Deke, J., Wei, T. & Kautz, T. *Asymdystopia: The threat of small biases in evaluations of education interventions that need to be powered to detect small impacts*. 1–62 (Institute of Education Sciences).
8. Allcott, H. Social norms and energy conservation. *J. Public Econ.* **95**, 1082–1095 (2011).
9. Kobes, A., Kretschmer, T., Timmerman, G. & Schreuder, P. Interventions aimed at preventing and reducing overweight/obesity among children and adolescents: a meta-synthesis. *Obes. Rev.* **19**, 1065–1079 (2018).
10. Stice, E., Shaw, H. & Marti, C. N. A meta-analytic review of obesity prevention programs for children and adolescents: the skinny on interventions that work. *Psychol. Bull.* **132**, 667–691 (2006).

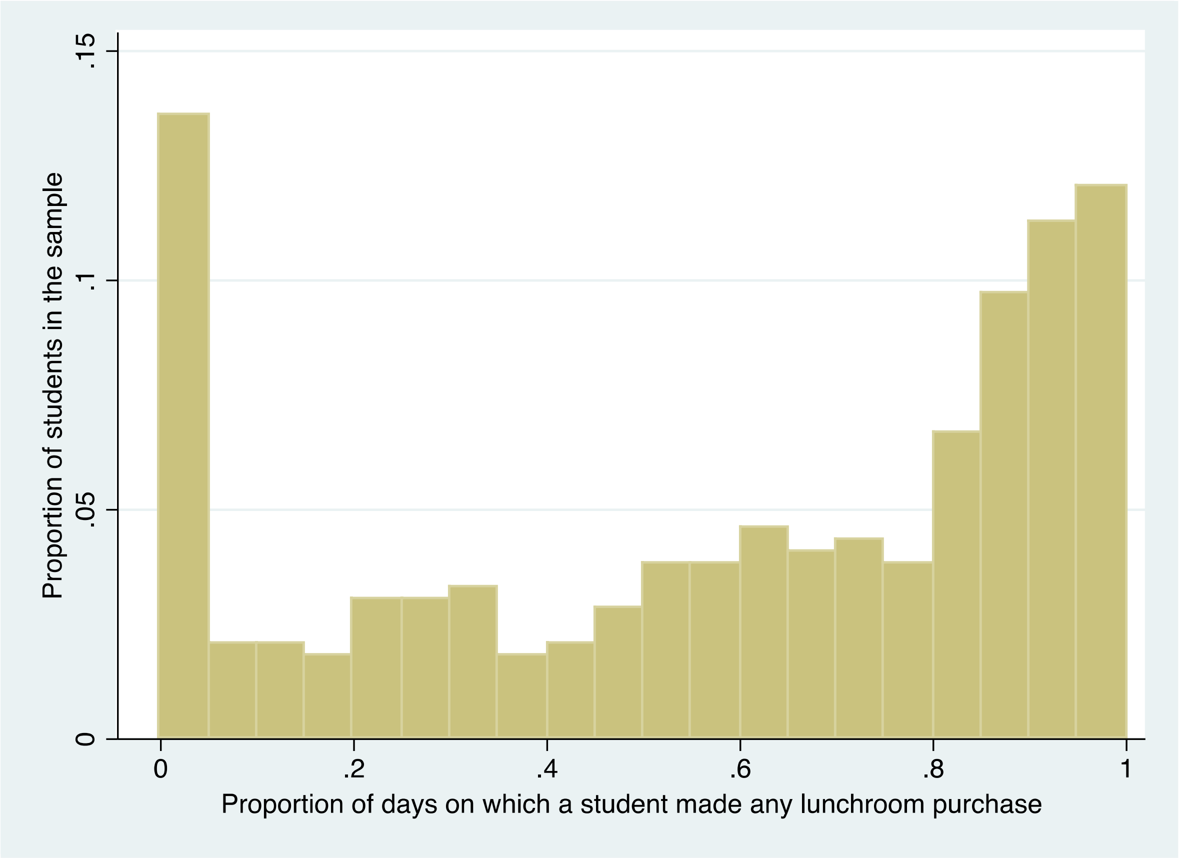
11. Simonsohn, U., Simmons, J. P. & Nelson, L. D. *Specification Curve: Descriptive and Inferential Statistics on All Reasonable Specifications*. (Social Science Research Network, 2015).

Supplementary Figures



Supplementary Figure 1

Effect of exposé intervention on (A) values-aligned construal of healthy eating, (B) social-status appeal of healthy eating, (C) reported anger in response to junk food advertising, and (D) reported desire to consume junk food products depicted in advertisements.



Supplementary Figure 2

Histogram depicting the proportion of students who made a cafeteria purchase using their student ID card on various proportions of school days over the course of the year.

Supplementary Table 1: Linear Mixed Effects Regressions Predicting Affect Misattribution Procedure (AMP) Responses

	Stimuli		
	All Stimuli (Healthy Reversed)	Junk Food Stimuli	Healthy Food Stimuli
	(1)	(2)	(3)
Exposé	-0.075	-0.089	0.055
	(0.022)	(0.030)	(0.025)
	P < 0.001	*P* = 0.004	*P* = 0.027
Junk	-0.116		
	(0.010)		
	P < 0.001		
2 Week	0.052	0.063	0.003
	(0.016)	(0.017)	(0.015)
	P < 0.001	*P* < 0.001	*P* = 0.83
3 Month	0.040	0.041	-0.003
	(0.016)	(0.017)	(0.015)
	P = 0.011	*P* = 0.017	*P* = 0.86
Exposé:Junk	0.013		
	(0.016)		
	P = 0.42		
Exposé:2 Week	0.018	0.024	-0.015
	(0.026)	(0.029)	(0.025)
	P = 0.49	*P* = 0.41	*P* = 0.56
2 Week:Junk	-0.045		
	(0.013)		
	P < 0.001		
Exposé:3 Month	0.031	0.045	-0.014
	(0.026)	(0.029)	(0.025)
	P = 0.22	*P* = 0.12	*P* = 0.59
3 Month:Junk	-0.032		
	(0.013)		
	P = 0.016		
Exposé:2 Week:Junk	-0.001		
	(0.022)		
	P = 0.98		
Exposé:3 Month:Junk	-0.011		
	(0.022)		
	P = 0.61		
Constant	0.485	0.594	0.694
	(0.015)	(0.022)	(0.018)
	P < 0.001	*P* < 0.001	*P* < 0.001
Observations	1,388	694	694

Supplementary Table 2: Linear Mixed Effects Regressions Predicting Lunchroom Choices, Full Sample

	Dependent Variable	
	Junk Snack	Healthy Snack
	(1)	(2)
Exposé group pre-intervention	0.009	0.002
	(0.014)	(0.002)
	P = 0.54	*P* = 0.42
Post-intervention (Pre = 0, Post = 1)	0.020	-0.001
	(0.006)	(0.001)
	P < 0.001	*P* = 0.67
Exposé:Post-intervention	-0.012	0.002
	(0.005)	(0.001)
	P = 0.027	*P* = 0.039
Constant	0.170	0.005
	(0.010)	(0.002)
	P < 0.001	*P* = 0.004
Observations	62,264	62,264

Supplementary Table 3: Linear Mixed Effects Regressions Predicting Lunchroom Choices, With Gender Interactions

	Dependent Variable			
	Junk Snack		Healthy Snack	
	(1)	(2)	(3)	(4)
Exposé group pre-intervention	0.023	-0.004	0.001	0.003
	(0.021)	(0.020)	(0.003)	(0.003)
	$P = 0.28$	$P = 0.83$	$P = 0.78$	$P = 0.35$
Post-intervention (Pre = 0, Post = 1)	0.033	0.015	-0.003	0.002
	(0.007)	(0.007)	(0.002)	(0.002)
	$P < 0.001$	$P = 0.034$	$P = 0.10$	$P = 0.24$
Female	0.041		-0.002	
	(0.022)		(0.003)	
	$P = 0.060$		$P = 0.64$	
Male		-0.041		0.002
		(0.022)		(0.003)
		$P = 0.060$		$P = 0.64$
Exposé:Post-intervention	-0.044	0.018	0.008	-0.003
	(0.008)	(0.008)	(0.002)	(0.002)
	$P < 0.001$	$P = 0.020$	$P < 0.001$	$P = 0.086$
Exposé:Female	-0.027		0.002	
	(0.029)		(0.004)	
	$P = 0.35$		$P = 0.65$	
Post-intervention:Female	-0.019		0.005	
	(0.008)		(0.002)	
	$P = 0.018$		$P = 0.011$	
Exposé:Post-intervention:Female	0.062		-0.011	
	(0.011)		(0.003)	
	$P < 0.001$		$P < 0.001$	
Exposé:Male		0.027		-0.002
		(0.029)		(0.004)
		$P = 0.35$		$P = 0.65$
Post-intervention:Male		0.019		-0.005
		(0.008)		(0.002)
		$P = 0.018$		$P = 0.011$
Exposé:Post-intervention:Male		-0.062		0.011
		(0.011)		(0.003)
		$P < 0.001$		$P < 0.001$
Constant	0.143	0.184	0.005	0.004
	(0.016)	(0.015)	(0.002)	(0.002)
	$P < 0.001$	$P < 0.001$	$P = 0.035$	$P = 0.12$
Observations	58,480	58,480	58,480	58,480

Supplementary Table 4: Linear Probability Models Predicting Dummy Variables Indicating Missing Data

	Missing Data Dummy Variable				
	Values-Alignment	Ad Perceptions	Snack Form	2wk AMP	3mo AMP
	(1)	(2)	(3)	(4)	(5)
Exposé	0.074	0.013	-0.057	-0.066	-0.126
	(0.085)	(0.132)	(0.138)	(0.129)	(0.098)
	P = 0.39	*P* = 0.92	*P* = 0.68	*P* = 0.61	*P* = 0.20
Age	0.050	0.065	0.026	-0.0004	0.032
	(0.018)	(0.027)	(0.029)	(0.027)	(0.020)
	P = 0.004	*P* = 0.017	*P* = 0.37	*P* = 0.99	*P* = 0.12
Male	0.055	0.038	0.029	-0.166	-0.061
	(0.036)	(0.056)	(0.059)	(0.055)	(0.042)
	P = 0.13	*P* = 0.50	*P* = 0.62	*P* = 0.003	*P* = 0.15
Disadv	0.022	-0.016	-0.051	-0.032	0.015
	(0.030)	(0.047)	(0.049)	(0.046)	(0.035)
	P = 0.46	*P* = 0.74	*P* = 0.30	*P* = 0.49	*P* = 0.66
White	-0.021	-0.034	-0.105	-0.042	-0.046
	(0.037)	(0.058)	(0.060)	(0.056)	(0.043)
	P = 0.56	*P* = 0.56	*P* = 0.083	*P* = 0.46	*P* = 0.28
Exposé:Age	-0.047	-0.096	-0.005	-0.044	-0.027
	(0.025)	(0.040)	(0.041)	(0.039)	(0.029)
	P = 0.067	*P* = 0.015	*P* = 0.91	*P* = 0.26	*P* = 0.36
Exposé:Male	-0.040	-0.035	-0.066	0.059	0.053
	(0.050)	(0.078)	(0.082)	(0.077)	(0.058)
	P = 0.43	*P* = 0.66	*P* = 0.42	*P* = 0.44	*P* = 0.36
Exposé:Disadv	-0.046	0.058	0.092	-0.040	0.020
	(0.042)	(0.065)	(0.068)	(0.064)	(0.049)
	P = 0.28	*P* = 0.38	*P* = 0.18	*P* = 0.53	*P* = 0.68
Exposé:White	-0.007	-0.011	0.100	0.005	0.017
	(0.052)	(0.081)	(0.085)	(0.080)	(0.061)
	P = 0.89	*P* = 0.89	*P* = 0.24	*P* = 0.95	*P* = 0.78
Constant	-0.026	0.125	0.220	0.402	0.174
	(0.062)	(0.097)	(0.102)	(0.095)	(0.072)
	P = 0.67	*P* = 0.20	*P* = 0.031	*P* < 0.001	*P* = 0.017
Observations	340	340	340	340	340
R^2	0.048	0.034	0.030	0.069	0.047
Adjusted R^2	0.017	0.001	-0.002	0.037	0.015

Supplementary Table 5: Linear Regression Models for Dependent Measures, with Demographic Interactions (Centred)

	Outcome				
	Values-Alignment	Ad Perceptions	Snack Form	2wk AMP	3mo AMP
	(1)	(2)	(3)	(4)	(5)
Exposé	1.007	0.974	0.028	-0.121	-0.096
	(0.095)	(0.115)	(0.094)	(0.038)	(0.036)
	P < 0.001	*P* < 0.001	*P* = 0.77	*P* = 0.002	*P* = 0.008
Age	-0.039	-0.073	0.090	-0.013	0.002
	(0.067)	(0.085)	(0.071)	(0.026)	(0.025)
	P = 0.56	*P* = 0.39	*P* = 0.21	*P* = 0.61	*P* = 0.94
Male	-0.178	-0.049	0.132	0.051	0.014
	(0.068)	(0.085)	(0.068)	(0.027)	(0.026)
	P = 0.01	*P* = 0.57	*P* = 0.053	*P* = 0.062	*P* = 0.60
Disadv	0.081	0.053	-0.018	0.007	-0.030
	(0.071)	(0.084)	(0.071)	(0.028)	(0.028)
	P = 0.26	*P* = 0.53	*P* = 0.80	*P* = 0.81	*P* = 0.27
White	0.046	0.043	-0.046	-0.008	-0.008
	(0.070)	(0.085)	(0.072)	(0.028)	(0.027)
	P = 0.51	*P* = 0.61	*P* = 0.52	*P* = 0.77	*P* = 0.76
Exposé:Age	-0.031	0.003	-0.153	-0.026	-0.065
	(0.098)	(0.120)	(0.097)	(0.038)	(0.036)
	P = 0.75	*P* = 0.98	*P* = 0.12	*P* = 0.50	*P* = 0.076
Exposé:Male	-0.044	-0.227	-0.133	-0.021	0.028
	(0.095)	(0.116)	(0.094)	(0.038)	(0.036)
	P = 0.64	*P* = 0.052	*P* = 0.16	*P* = 0.57	*P* = 0.44
Exposé:Disadv	-0.231	0.010	-0.021	0.018	0.057
	(0.099)	(0.119)	(0.097)	(0.039)	(0.038)
	P = 0.02	*P* = 0.93	*P* = 0.83	*P* = 0.65	*P* = 0.13
Exposé:White	0.017	-0.054	-0.016	-0.002	0.005
	(0.099)	(0.119)	(0.098)	(0.040)	(0.038)
	P = 0.86	*P* = 0.65	*P* = 0.87	*P* = 0.96	*P* = 0.90
Constant	2.560	1.264	2.344	-0.073	-0.081
	(0.073)	(0.089)	(0.074)	(0.029)	(0.028)
	P < 0.001	*P* < 0.001	*P* < 0.001	*P* = 0.012	*P* = 0.004
Observations	321	290	284	290	314
R^2	0.310	0.243	0.031	0.067	0.065
Adjusted R^2	0.285	0.213	-0.008	0.031	0.031