Examining Consumer Responses to Restaurant Menu Labeling Requirements

OMB No. OS-0990-XXXX

Supporting Statement – Section B

Submitted: October 13, 2015, revised March 16, 2016, revised April 12, 2016

Program Official/Project Officer
Amber Jessup, Ph.D.
Senior Economist
U.S. Department of Health and Human Services
Office of the Assistant Secretary for Planning and Evaluation
200 Independence Avenue SW, Washington DC 20201
(202) 690-6621
amber.jessup@hhs.gov

SUPPORTING STATEMENT

EXAMINING CONSUMER RESPONSES TO RESTAURANT MENU LABELING REQUIREMENTS

B. Collection of Information Employing Statistical Methods

1. Respondent Universe and Respondent Selection

This project uses the American Life Panel (ALP) for this project. The American Life Panel (ALP) is a high quality sample with national representativeness. While ALP has been used for national estimates, national representativeness is not required in this study, whose primary aims is how people respond to menu labeling in different settings. Nevertheless, external validity is increased by sampling from a nationally representative sample than a convenience sample as for most internet panel

We will use a random sample of ALP participants (see below for exact numbers). There is no oversampling.

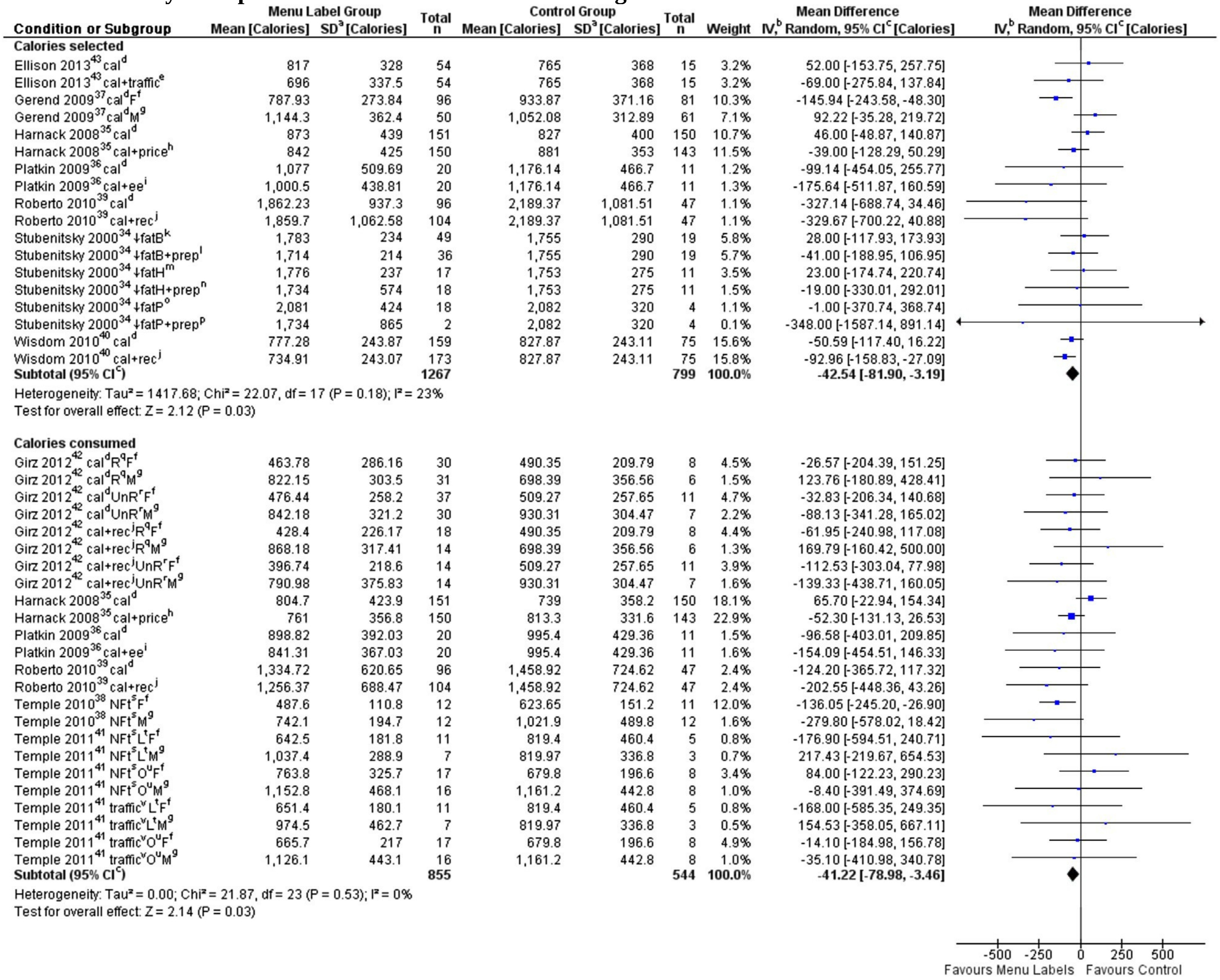
ALP recruits participants from several sources, including the University of Michigan Monthly Survey, the National Survey Project cohort, and several targeted recruitment methods to add specific populations (e.g. active recruitment for vulnerable populations). Such recruitment methods include address-based sampling. Computer ownership or Internet access was not a requirement for ALP in order to eliminate the bias found in other Internet survey panels. For individuals without their own internet access, RAND provides panel members with Internet access by providing a WebTV and an Internet subscription.

Power Calculation

The sample size (2000 completed responses) was determined by the budget. Nevertheless, we conducted power calculations to confirm that this would be sufficient to detect meaningful effects of menu labeling, based on a review of sample sizes and variances in prior field and experimental studies. Power calculations are inherently speculative and no previous work can directly inform this experiment. The best source is a review of quasi-experimental and actual experimental studies (Sinclair et al., 2014). Our best guess of a mean population effect of labeling calories with contextual information across a variety of settings is 67.

Next, we need to get an estimate about likely variances or standard deviation in responses. The table below summarizes previous experiments. The last column shows the confidence intervals and also indicates the problem with sample sizes in prior studies: Confidence intervals are very wide in studies with 30-300 participants total. Wisdom et al. is an outlier with unusually small confidence intervals despite the small sample.

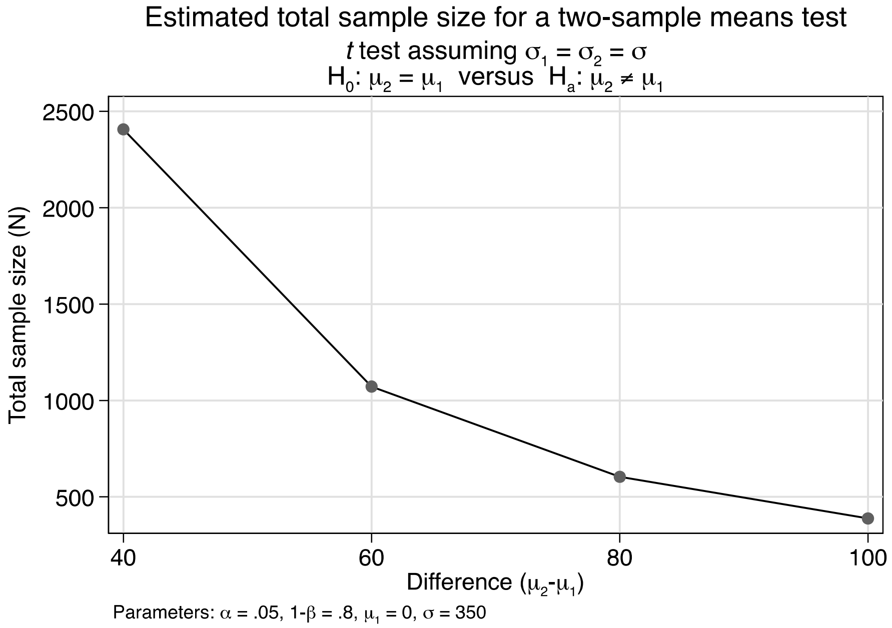
Table: Summary of Experimental Studies on Menu Labeling



	Menu Label Group		Total	Control Group		Total		Mean Difference
Condition or Subgroup	Mean [Calories]	SD[a] [Calories]	n	Mean [Calories]	SD[a] [Calories]	n	Weight	IV,[b] Random, 95% CI[c] [Calories]
Calories selected								
Ellison 2013[43] cal[d]	817	328	54	765	368	15	3.2%	52.00 [-153.75, 257.75]
Ellison 2013[43] cal+traffic[e]	696	337.5	54	765	368	15	3.2%	-69.00 [-275.84, 137.84]
Gerend 2009[37] cal[d]F[f]	787.93	273.84	96	933.87	371.16	81	10.3%	-145.94 [-243.58, -48.30]
Gerend 2009[37] cal[d]M[g]	1,144.3	362.4	50	1,052.08	312.89	61	7.1%	92.22 [-35.28, 219.72]
Harnack 2008[35] cal[d]	873	439	151	827	400	150	10.7%	46.00 [-48.87, 140.87]
Harnack 2008[35] cal+price[h]	842	425	150	881	353	143	11.5%	-39.00 [-128.29, 50.29]
Platkin 2009[36] cal[d]	1,077	509.69	20	1,176.14	466.7	11	1.2%	-99.14 [-454.05, 255.77]
Platkin 2009[36] cal+ee[i]	1,000.5	438.81	20	1,176.14	466.7	11	1.3%	-175.64 [-511.87, 160.59]
Roberto 2010[39] cal[d]	1,862.23	937.3	96	2,189.37	1,081.51	47	1.1%	-327.14 [-688.74, 34.46]
Roberto 2010[39] cal+rec[j]	1,859.7	1,062.58	104	2,189.37	1,081.51	47	1.1%	-329.67 [-700.22, 40.88]
Stubenitsky 2000[34] ↓fatB[k]	1,783	234	49	1,755	290	19	5.8%	28.00 [-117.93, 173.93]
Stubenitsky 2000[34] ↓fatB+prep[l]	1,714	214	36	1,755	290	19	5.7%	-41.00 [-188.95, 106.95]
Stubenitsky 2000[34] ↓fatH[m]	1,776	237	17	1,753	275	11	3.5%	23.00 [-174.74, 220.74]
Stubenitsky 2000[34] ↓fatH+prep[n]	1,734	574	18	1,753	275	11	1.5%	-19.00 [-330.01, 292.01]
Stubenitsky 2000[34] ↓fatP[o]	2,081	424	18	2,082	320	4	1.1%	-1.00 [-370.74, 368.74]
Stubenitsky 2000[34] ↓fatP+prep[p]	1,734	865	2	2,082	320	4	0.1%	-348.00 [-1587.14, 891.14]
Wisdom 2010[40] cal[d]	777.28	243.87	159	827.87	243.11	75	15.6%	-50.59 [-117.40, 16.22]
Wisdom 2010[40] cal+rec[j]	734.91	243.07	173	827.87	243.11	75	15.8%	-92.96 [-158.83, -27.09]
Subtotal (95% CI[c])			**1267**			**799**	**100.0%**	**-42.54 [-81.90, -3.19]**
Heterogeneity: Tau² = 1417.68; Chi² = 22.07, df = 17 (P = 0.18); I² = 23%								
Test for overall effect: Z = 2.12 (P = 0.03)								
Calories consumed								
Girz 2012[42] cal[d]R[q]F[f]	463.78	286.16	30	490.35	209.79	8	4.5%	-26.57 [-204.39, 151.25]
Girz 2012[42] cal[d]R[q]M[g]	822.15	303.5	31	698.39	356.56	6	1.5%	123.76 [-180.89, 428.41]
Girz 2012[42] cal[d]UnR[r]F[f]	476.44	258.2	37	509.27	257.65	11	4.7%	-32.83 [-206.34, 140.68]
Girz 2012[42] cal[d]UnR[r]M[g]	842.18	321.2	30	930.31	304.47	7	2.2%	-88.13 [-341.28, 165.02]
Girz 2012[42] cal+rec[j]R[q]F[f]	428.4	226.17	18	490.35	209.79	8	4.4%	-61.95 [-240.98, 117.08]
Girz 2012[42] cal+rec[j]R[q]M[g]	868.18	317.41	14	698.39	356.56	6	1.3%	169.79 [-160.42, 500.00]
Girz 2012[42] cal+rec[j]UnR[r]F[f]	396.74	218.6	14	509.27	257.65	11	3.9%	-112.53 [-303.04, 77.98]
Girz 2012[42] cal+rec[j]UnR[r]M[g]	790.98	375.83	14	930.31	304.47	7	1.6%	-139.33 [-438.71, 160.05]
Harnack 2008[35] cal[d]	804.7	423.9	151	739	358.2	150	18.1%	65.70 [-22.94, 154.34]
Harnack 2008[35] cal+price[h]	761	356.8	150	813.3	331.6	143	22.9%	-52.30 [-131.13, 26.53]
Platkin 2009[36] cal[d]	898.82	392.03	20	995.4	429.36	11	1.5%	-96.58 [-403.01, 209.85]
Platkin 2009[36] cal+ee[i]	841.31	367.03	20	995.4	429.36	11	1.6%	-154.09 [-454.51, 146.33]
Roberto 2010[39] cal[d]	1,334.72	620.65	96	1,458.92	724.62	47	2.4%	-124.20 [-365.72, 117.32]
Roberto 2010[39] cal+rec[j]	1,256.37	688.47	104	1,458.92	724.62	47	2.4%	-202.55 [-448.36, 43.26]
Temple 2010[38] NFt[s]F[f]	487.6	110.8	12	623.65	151.2	11	12.0%	-136.05 [-245.20, -26.90]
Temple 2010[38] NFt[s]M[g]	742.1	194.7	12	1,021.9	489.8	12	1.6%	-279.80 [-578.02, 18.42]
Temple 2011[41] NFt[s]L[t]F[f]	642.5	181.8	11	819.4	460.4	5	0.8%	-176.90 [-594.51, 240.71]
Temple 2011[41] NFt[s]L[t]M[g]	1,037.4	288.9	7	819.97	336.8	3	0.7%	217.43 [-219.67, 654.53]
Temple 2011[41] NFt[s]O[u]F[f]	763.8	325.7	17	679.8	196.6	8	3.4%	84.00 [-122.23, 290.23]
Temple 2011[41] NFt[s]O[u]M[g]	1,152.8	468.1	16	1,161.2	442.8	8	1.0%	-8.40 [-391.49, 374.69]
Temple 2011[41] traffic[v]L[t]F[f]	651.4	180.1	11	819.4	460.4	5	0.8%	-168.00 [-585.35, 249.35]
Temple 2011[41] traffic[v]L[t]M[g]	974.5	462.7	7	819.97	336.8	3	0.5%	154.53 [-358.05, 667.11]
Temple 2011[41] traffic[v]O[u]F[f]	665.7	217	17	679.8	196.6	8	4.9%	-14.10 [-184.98, 156.78]
Temple 2011[41] traffic[v]O[u]M[g]	1,126.1	443.1	16	1,161.2	442.8	8	1.0%	-35.10 [-410.98, 340.78]
Subtotal (95% CI[c])			**855**			**544**	**100.0%**	**-41.22 [-78.98, -3.46]**
Heterogeneity: Tau² = 0.00; Chi² = 21.87, df = 23 (P = 0.53); I² = 0%								
Test for overall effect: Z = 2.14 (P = 0.03)								

Our first calculation shows what sample sizes are needed for detecting various effects with acceptable power, using a two-sample comparison (Figure 1). We use a standard deviation of 350 calories, about the pooled value in Ellison et al. (2013) or Gerend et al. (2009), which is smaller than the studies by Harnack and Platkin, but larger than Stubenitsky and Wisdom. The calculations are for a two-sided two-sample test with alpha=0.05 and 80% power. N is the combined sample size, so 1000 would be for two groups of 500 each and would have good power to detect the mean effect of 67 kcals. That means we would have good statistical power for comparing 4 subgroups within each setting. Probably we have better power for comparing each of those subgroups across settings, which would be a pairwise comparison as the same individual is considered twice and much of the variation in food choices is between rather than within individuals.

Figure 1: Total sample sizes needed for 80% power as a function of effect size[1]



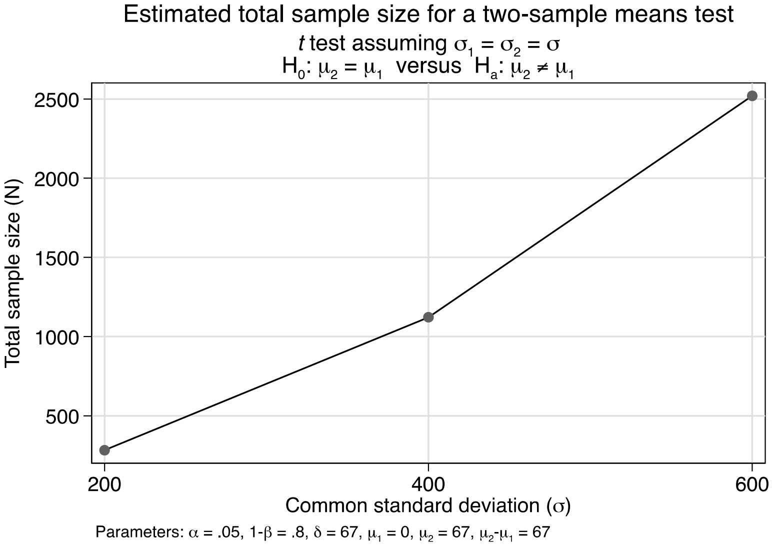
We expect that the standard deviation depends highly on the range of menu options and increases with a broader range of options and decreases with a smaller range. A setting like Starbucks would see smaller variances than Outback Steakhouse. However, we have no data that would allow us to relate the variance in calories in the menu to the variance in calories of choices. So our next calculation is about the influence of variances. We assume the true mean effect is 67 cals and calculate sample sizes when standard deviation across settings range from 200 to 600 kcals. We can see that for settings with very low variation, subsample analysis for small subgroups will be feasible. However, in settings were the standard deviations in individual choices reaches 600, even our full data set will be insufficient to detect the mean effect of 67 (with acceptable power).

Those calculations are only illustrative as our main analytic approach is a regression analysis rather than stratification. There are two countervailing effects: Stratification or

[1] Code: power twomeans 0 , sd(350) power(0.8) diff (40 60 80 100) graph

additional parameters reduce statistical power everything else being equal (in particular, the variance within a subgroup equals the population variance). However, regression models or stratification typically also reduce the residual variance as subgroups are more homogeneous, thus increasing statistical power. No data from prior studies exist to assess the relative magnitude.

Figure 2: Total sample size needed to detect an effect of 67 cals as a function of standard deviation in calorie choices[2]

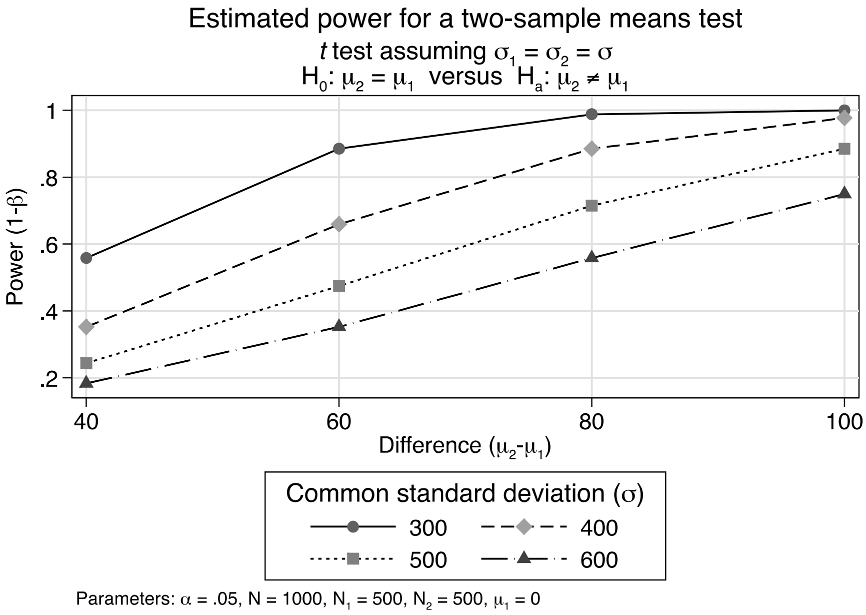


Finally, for our preferred sample size (500 per group), we calculate how standard deviations in calorie choice and effect sizes affect the statistical power.

Figure 3: Statistical power as a function of standard deviations and effect sizes. [3]

[2] power twomeans 0 , sd(200 400 600) power(0.8) diff (67) graph

[3] power twomeans 0 , sd(300 400 500 600) n(1000) diff (40 60 80 100) graph



2. Data Collection Procedures

The survey will be programmed and fielded using MMIC software (Multimode Interviewing Capability) on RAND's American Life Panel (ALP). The 20 minute surveys will provide more data than a 30 minute or even longer survey on a newly recruited sample because baseline sociodemographics have been collected for this panel and we do no have to ask those questions. The ALP website can be found here: https://mmicdata.rand.org/alp/

Once OMB approval is received, RAND will program the final instrument for administration in the American Life Panel system using MMIC software and then pre-test it on a small sample of up to 100 participants from ALP. The pretest will be concluded within 8 weeks of OMB approval.

ALP creates an analytic data file to which RAND will merge relevant information from previous data collections, including demographics and variables like self-reported height and weight, using the MMIC data management system. While simple tests of means, possibly stratified by subgroups, would provide unbiased and internally valid results (it is an randomized experiments), our primary approach will be regression analysis to estimate how menu labeling affects calorie choices in different settings across different settings by sociodemographics. Additional statistical models may be used to analyze discrete choices, from standard economic models (such as multinomial or nested multinomial models) to models incorporating possible violations of classic economic models (e.g. attribute-non-attention).

Although national representativeness is not a requirement for the study question (differential effects to menu labeling by type of restaurant setting and sociodemographics), it enhances external validity. We do not plan to weight regression models, although we would use weights for descriptive statistics. As with all surveys based on random samples, the composition of the un-weighted sample will differ from the population composition. RAND constructs sampling weights to correct for this sampling error and to make a weighted sample representative of US population, benchmarking it against the Current Population Survey (CPS). This choice follows common practice in surveys of consumers, for example, the Health and Retirement Study (HRS). Raking was found to give the best results as it allows finer categorizations of variables of interest (in particular, age) than cell-based post-stratification does, while still matching these distributions exactly. Variables were created that account for interactions with gender or with the number of household members, as described below, so that distributions are matched separately for males and females, and for number of household members. Specifically, the following distributions are matched exactly is:
Gender x Age, with 10 Categories:
male, 18-32
male, 33-43
male, 44-54
male, 55-64
male, 65+
Categories (6)-(10) are the same as (1)-(5), except that they are for females instead of males.
Gender x age, with 10 categories: (1) male, 18-32; (2) male, 33-43; (3) male, 44-54; (4) male, 55-64; (5) male, 65+. Categories (6)-(10) are the same as (1)-(5), except that they are for females instead of males.
Gender x race/ethnicity, with 6 categories: (1) male, non-Hispanic white; (2) male, non-Hispanic African American; (3) male, Hispanic and other; (4) female, non-Hispanic white; (5) female, non-Hispanic African American; (6) female, Hispanic and other.
Gender x education, with six categories: (1) male, high school or less; (2) male, some college or associate's degree; (3) male, bachelor's degree or more; (4) female, high school or less; (5) female, some college or associate's degree; (6) female, bachelor's degree or more. All aggregate U.S. statistics for the SCPC were weighted using the sampling weights constructed in this manner.
Number of household members x (household) income, with twelve categories: (1) household with one individual, <$25,000; (2) household with one individual, $25,000-$49,999; (3) household with one individual, $50,000-$74,999; (4) household with one individual, $75,000+. Categories (5)-(8) are the same as (1)-(4), but for households with two individuals. Categories (9)-(12) are the same as (1)-(4), but for households with more than two individuals.

The Figures below show how the weighted ALP data compares to the US estimates from the CPS:

Figures

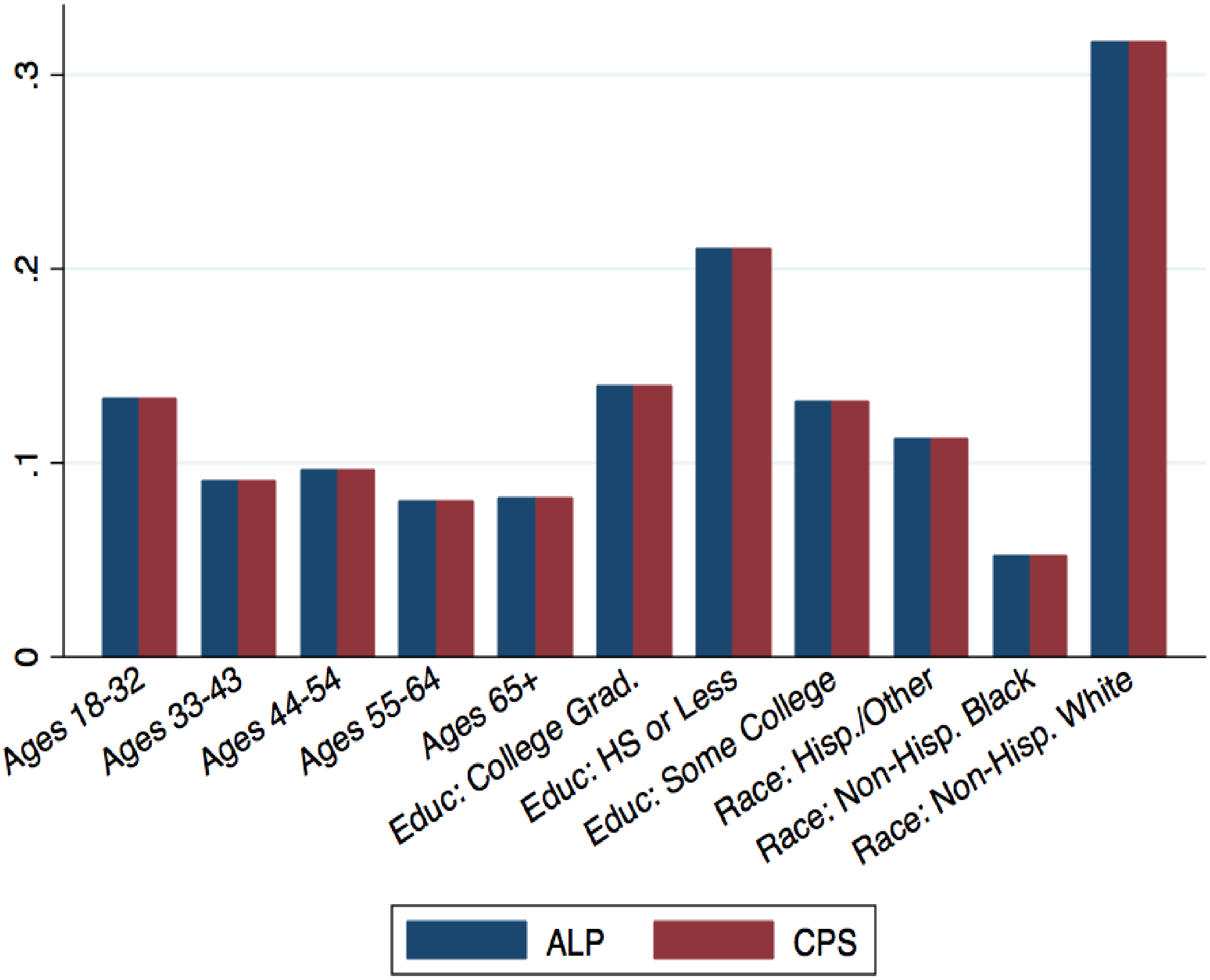


Figure 1: Comparison of weighted frequencies in ALP and CPS, Males

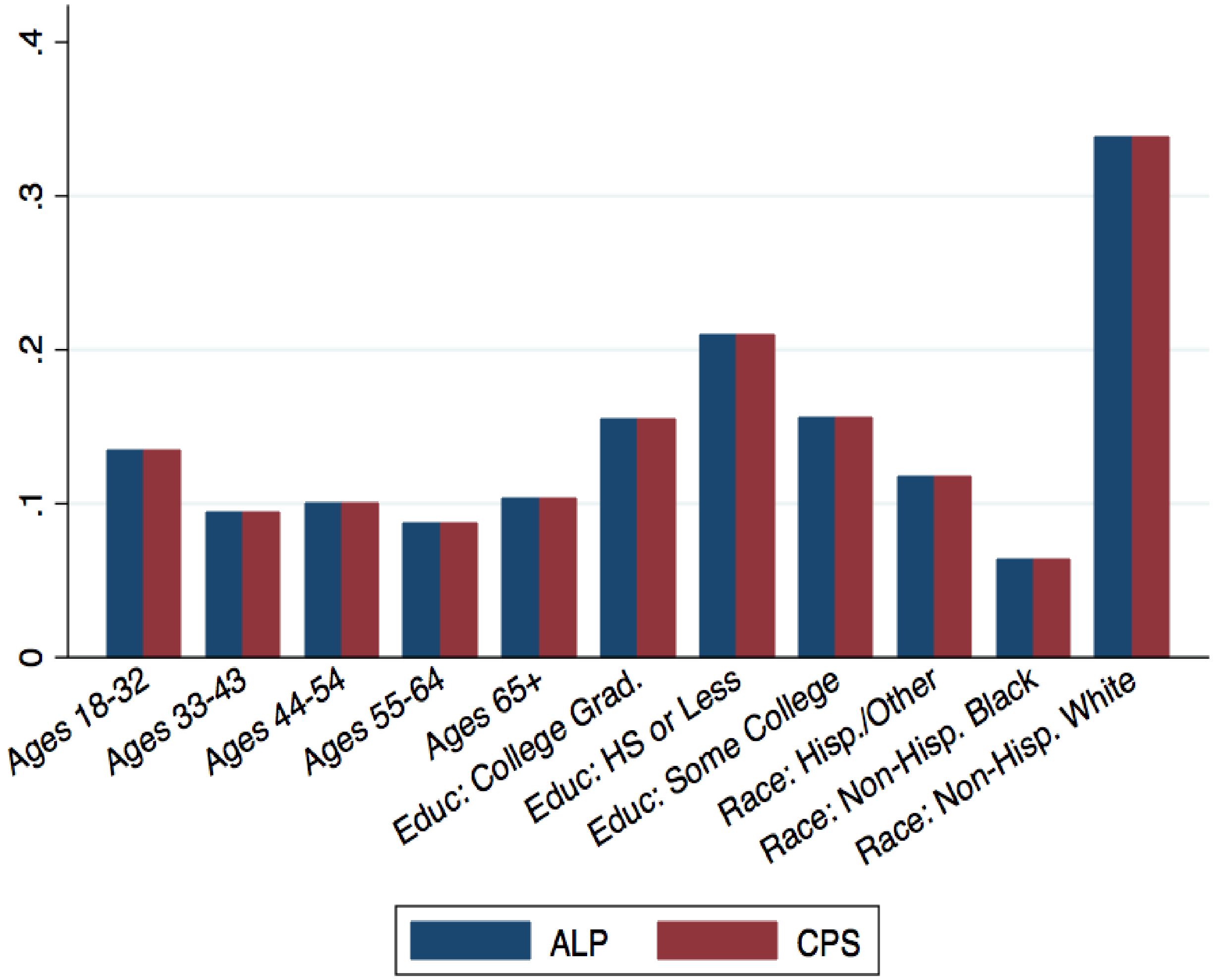


Figure 2: Comparison of weighted frequencies in ALP and CPS, Females

Respondents of the survey are randomly assigned to different menus with and without calorie labeling. The primary goal is to estimate how calorie labeling differentially affect choices in different type of food outlets and consumers with different individual characteristics. Secondary goals are to estimate how consumers trade off prices and calories, and to calculate the welfare gains from labeling.

Each respondent will be presented with nine different menus (see list in Table 1). He/she will be asked to make food choices from each menu, followed by the final section of the survey where respondents will answer attitudinal and behavioral questions. The latter include questions about how hungry the respondent is at the time of the survey, how important characteristics like low price, value portions, and low calories are, and how much they generally pay attention to calorie and nutritional information.

Table 1: The 9 types of food outlets

Fast food burger chains
Fast casual Asian restaurants
Ice cream parlors
Movie theatre snack bars
Pizza-by-the-slice stands
Organic, locally sourced restaurants
Fast casual Mexican restaurants
Salad/sandwich restaurants
Coffee shops

For each respondent, we will randomize the survey in the following ways:
1. For each individual the order of food outlets (fast food, Asian, ice cream, etc) will be randomly assigned. This will prevent any potential bias due to the order of appearance of the menus/food outlets.

2. For each food outlet the labeling of the menu shown is also randomly determined. This is the primary experiment, with the following treatments:

- Treatment A: no calorie labels (this will serve as the "control")
- Treatment B: with calorie labels which meet the requirements of the FDA's new regulation, i.e. the size of the calorie declaration must be no smaller than the size of the name or the price of the menu item it refers to, whichever is smaller. In general such calorie declarations must be in the same color, or a similar color as that used for the name of the associated menu item. The contextual statement about recommended daily caloric intake is shown. This is the "do minimum" treatment in which the new regulation will be met just barely.
- Treatment C (for only four of the food outlets): We allow an alternative labeling design for the fast casual Asian restaurant, the salad/sandwich restaurant, the pizza-by-the-slice stand and the organic, locally sourced restaurant. The design will meet the requirements of the new regulation and use fonts that are more pronounced than Treatment B (e.g. through the use of a heavier font and/or colors that stands out from the background). While many restaurants will use minimal requirements, some are likely to feature calories more prominently (as Subway has done for a long time). This design will allow separating visibility from other restaurant effects (e.g. intentional health halo).

Because we have nine food outlets, five of them have two treatments (A and B), and four of them have three treatments (A to C), it is not possible to ensure that each individual will be shown equal numbers of Treatment A, B, and C. The extent that this may or may not introduce respondent bias will be determined empirically.

3. For some food outlets (fast-food burger, ice-cream, movie theatre, fast casual Mexican and coffee), the menus shown with have varying sets of prices. The options are:

- Default prices
- Lower calorie choices are approximately 20% cheaper (a "healthy dining subsidy")
- High calorie choices are approximately 20% more expensive (a "fat tax").

This price manipulation breaks the perfect collinearity between prices and calories and allows the study team to estimate the price sensitivity and eventually the consumer gain from better choices.

3. Expected Response Rates and Methods to Assure Optimize Response Rates

For this data collection, we will target 2,000 completed responses of a 20 minute survey. Based on recent ALP surveys, we expect a 70% response rate. There have been over 430 surveys fielded using the ALP and the average response rate has been 70%. This is an average, so some surveys have done better and some worse, but we are expecting to get about 70%. Most surveys fielded to the ALP are about 20-25 minutes long and respondents are usually offered an incentive of $10-$20. Therefore, we believe our survey should fall within the average. The response rate does depend somewhat on how long the survey remains in the field and how many reminders are sent. We plan to field the survey for 2 to 4 weeks, depending on how quickly we reach our targeted response rate. Therefore, we will invite 2,850 individuals to participate in order to reach 2000 completes (we use reminders and incentives to achieve at least this response and also roll out samples in waves to assure our target completion rates). In previous ALP surveys, most individuals completing the interview respond within one week of the date the survey went into the field. In addition, through the MMIC system, we can send customized email reminders once per week for up to 4 weeks to panel members who have started the survey but not completed it and to those who have not started it. The reminders combined with the incentive are used to get to the 70% response rate.

4. Tests of Procedures or Methods

The RAND team has already heavily tested the survey to ensure the timing of the survey as well as to ensure that there are no problems with the MMIC programming of the tool or the wording of the questions. To test within the RAND team, we used eight graduate students from the Pardee RAND graduate school and did two focus groups with them. In each, the students were first instructed to go through the entire survey to test the length (average times were 21 minutes for the first group and 18 minutes for the second with no major outliers). Once that test was completed, two members of the RAND team lead discussions with the group to look at the wording of the questions and ensure that there was no confusion. The testers did not find any major problems with the survey. There were no extreme outliers in the timing either. These testers are likely to have more education than the average ALP member, so once OMB approval is received, the survey will be fielded to a small part of the sample (about 100 members of the ALP) as the first wave to ensure that there are no issues with the survey itself and that the timing remains at or below an average of 20 minutes per survey. However, the graduate student testers all regularly order food from the types of restaurants in the survey.

5. Statistical and Data Collection Consultants

The survey, sampling approach, and data collection procedures were designed by the RAND Corporation under the leadership of:

Roland Sturm, Ph.D.
RAND Corporation
1776 Main Street
Santa Monica, CA 90407