Experimental Studies of Cigarette Graphic Warning Labels: Analysis Plan

1. Respondent Universe and Sampling Methods

The respondent universe for this study is (1) adolescent current cigarette smokers aged 13-17 years old; (2) adolescents who are susceptible to initiation of cigarette smoking aged 13-17 years old; (3) young adult current cigarette smokers aged 18 to 24 years old; (4) young adult nonsmokers aged 18 to 24 years old; (5) older adult current cigarette smokers aged 25 years old and older; and (6) older adult nonsmokers aged 25 years old and older.

Study participants will be recruited from a national online panel of adults managed by Lightspeed. The Lightspeed panel is a non-probability convenience sample recruited via social media, online recruitment (e.g. via banner placements), and affiliate corporate networks. For the current study, Lightspeed will recruit adult panelists and parents of potential youth respondents using information from panelist's user profiles related to study eligibility (i.e. age, smoking status, and whether or not the panelist has a child in the eligible age range).

Lightspeed panel members will receive an email inviting them to participate in the study. Adolescent children of adult panel participants will be invited to complete the survey through an email invitation to their parents asking for their consent to solicit their child's opinions. Panel members and children of panelists who choose to participate will complete the questionnaire. During the study recruitment, we will monitor the sample and can adjust recruitment targeting as needed to ensure the sociodemographic distribution is diverse in terms of age, gender, education, and ethnicity/race. We estimate a total of 9,760 respondents will complete the baseline data collection. We will administer two follow-up surveys of baseline survey participants, with estimated retention rates of between 50% - 100% for each follow-up session (Table 1).

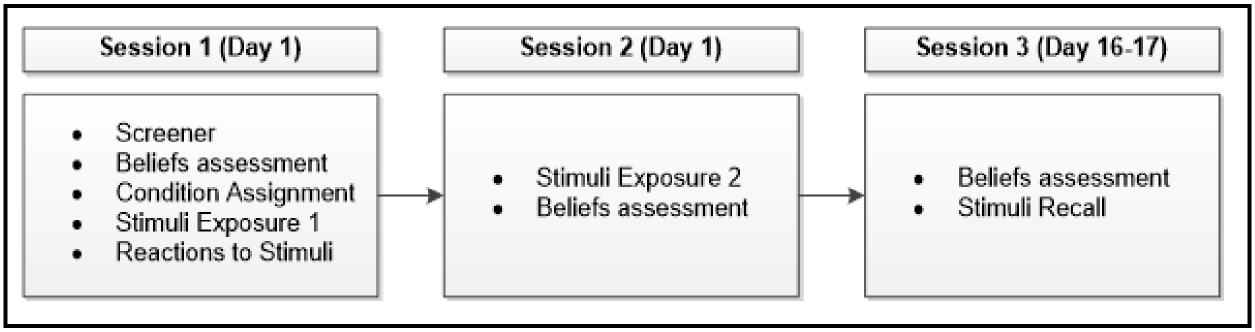
Table 1. Estimated Sample Sizes, by Sample Group and Study Session

Age Group	Smoking Status	Session		
		1	2	3
Adolescent	Current Smoker	230	115 - 230	58 - 230
Adolescent	Susceptible to Smoking	2,070	1,035 - 2,070	518 – 2,070
Young Adult	Current Smoker	1,330	665 – 1,330	333 – 1,330
Young Adult	Nonsmoker	1,330	665 – 1,330	333 – 1,330
Older Adult	Current Smoker	2,400	1,200 – 2,400	600 – 2,400
Older Adult	Nonsmoker	2,400	1,200 – 2,400	600 – 2,400
Total		9,760	4,880 – 9,760	2,440 – 9,760

2. Study Protocol

The study comprises three Sessions, outlined in Figure 1.

Figure 1. Study Protocol



In Session 1, participants will first complete a screening questionnaire through an email invitation. After screening for inclusion (see Study Screener), participants who qualify for the study will complete three consecutive components: (1) a baseline assessment of beliefs about the negative health consequences of cigarette smoking; (2) assignment to study condition and exposure to cigarette warning stimuli according to condition assignment; and (3) assessment of new information, self-reported learning, and other reactions to the stimuli (see Session 1 Survey Instrument). These three components are described below.

- *Component (1):* First, participants will be asked questions about beliefs related to the health consequences of cigarette smoking.
- *Component (2):* Following the baseline assessment of beliefs, participants will be randomized to one of 16 treatment conditions or a control condition with variation in exposure to cigarette warnings (Table 2 illustrates the targeted Session 1 sample size and allocation across experimental groups, within each age group and overall). Participants in each experimental condition will be exposed to one graphic health warning (GHW), with each condition corresponding to a unique warning from a set of 16. Participants in the control condition will be exposed to a random selection of one of four Surgeon General's (SG) warnings. Each stimuli exposure will include viewing of the warning in two formats: on a mock cigarette package depicted in a 3-dimensional, rotational model; and on a mock cigarette advertisement. In all analyses, stimuli exposure will be considered the joint exposure to both stimuli formats; stimuli format is not considered a study factor
- *Component (3):* After viewing the warning stimuli in both package and advertisement formats, participants will complete a brief set of measures to assess (a) if the information presented in the warning was new; (b) self-reported learning from the warning; (c) if the warning was easy to understand; (d) if the warning was perceived to be a fact or an

opinion; (e) if the warning was informative; (f) if the warning grabbed their attention; and (g) if the warning made them think about the health risks of smoking.

One to two days following completion of the baseline assessment (Session 1), participants will receive an email invitation to complete a follow-up (Session 2). In this follow-up session, participants will first be re-exposed to the warning stimuli they were shown in Session 1. This exposure will follow the same protocol described in Component 2, above. Following stimuli exposure, participants will complete a set of immediate post-test measures assessing beliefs related to the negative health consequences of cigarette smoking (see Session 2 Survey Instrument). These measures will facilitate an assessment of change in beliefs related to smoking-related health consequences following exposure to the cigarette warning stimuli.

Approximately 14 days later, at the delayed post-test (Session 3), participants will receive an email invitation to complete a questionnaire assessing measures of beliefs about the negative health consequences of cigarette smoking, as well as recall of the warning (see Session 3 Survey Instrument).

Table 2. Condition Assignment and Session 1 Sample Size, by Study Population

Condition	Warning Type	Exposure	Expected Sample Size at Session 1			
			Adolescents	Young Adults	Older Adults	Total
0 (Control)	SG	Random selection of 1 of the following SG warnings: 1) SURGEON GENERAL'S WARNING: Smoking Causes Lung Cancer, Heart Disease, Emphysema, and May Complicate Pregnancy. 2) SURGEON GENERAL'S WARNING: Quitting Smoking Now Greatly Reduces Serious Risks to Your Health. 3) SURGEON GENERAL'S WARNING: Smoking by Pregnant Women May Result in Fetal Injury, Premature Birth, and Low Birth Weight. 4) SURGEON GENERAL'S WARNING: Cigarette Smoke Contains Carbon Monoxide.	492	564	1,024	2,080
1	GHW	WARNING: Cigarettes are addictive.	113	131	236	480
2	GHW	WARNING: Tobacco smoke can harm your children.	113	131	236	480
3	GHW	WARNING: Smoking can kill you.	113	131	236	480
4	GHW	WARNING: Tobacco smoke causes fatal lung disease in nonsmokers.	113	131	236	480
5	GHW	WARNING: Quitting smoking now greatly reduces serious risks to your health.	113	131	236	480
6	GHW	WARNING: Smoking causes head and neck cancer.	113	131	236	480

7	GHW	WARNING: Smoking causes bladder cancer, which can lead to bloody urine.	113	131	236	480
8	GHW	WARNING: Smoking during pregnancy stunts fetal growth.	113	131	236	480
9	GHW	WARNING: Smoking can cause heart disease and strokes by clogging arteries.	113	131	236	480
10	GHW	WARNING: Smoking causes COPD, a lung disease that can be fatal. [IMAGE 1: BRAIN]	113	131	236	480
11	GHW	WARNING: Smoking causes COPD, a lung disease that can be fatal. [IMAGE 2: MAN]	113	131	236	480
12	GHW	WARNING: Smoking reduces blood flow, which can cause erectile dysfunction.	113	131	236	480
13	GHW	WARNING: Smoking reduces blood flow to the limbs, which can require amputation.	113	131	236	480
14	GHW	WARNING: Smoking causes type 2 diabetes, which raises blood sugar.	113	131	236	480
15	GHW	WARNING: Smoking causes age-related macular degeneration, which can lead to blindness.	113	131	236	480
16	GHW	WARNING: Smoking causes cataracts, which can lead to blindness.	113	131	236	480
TOTAL			**2,300**	**2,660**	**4,800**	**9,760**

NOTE: SG = Surgeon General’s Warning; GWH = Graphic Health Warning

3. Power Analysis

We conducted power calculations to confirm that the overall sample size (shown in Table 2) is sufficiently powered and to determine the optimal sample size and allocation of sample across study conditions. To control for Type 1 error taking into account multiple testing, power calculations were based on the false discovery rate (FDR) (Benjamini & Hochberg, 1995). Assuming the tests are independent, the FDR is the expected proportion of significant results that are falsely declared as statistically significant. Controlling the FDR is controlling the expected proportion of falsely declared differences (i.e., false discoveries). Controlling the FDR is a more powerful method for dealing with multiple comparisons than other methods that control the family-wise error rate (Benjamini & Hochberg, 1995).

For the overall study sample size, we calculated power to detect a difference in the change in a tobacco-related belief from Session 1 to Session 2 between treatment and control groups (i.e., difference in difference) (Table 3 provides power estimates for Session 2 across various scenarios). RTI calculated power to detect a 0.3 difference on a 7-point scale (two-sided tests, assuming a standard deviation of 1) under different scenarios with variation in FDR, within-person correlation between Session 1 and 2, and sample allocation. RTI conservatively assumed 50% retention from Session 1 to Session 2. Power calculations were computed using 100 simulations for each sample allocation in SAS v9.4.

Across various assumptions of within-person correlation and FDR, we found generally higher levels of power using an optimized sample allocation with between 1,760 and 2,400 participants assigned to the control condition at Session 1 (880-1,200 participants at Session 2, assuming 50% retention). Based on this analysis showing that higher power is achieved with an unbalanced allocation, FDA plans to allocate 2,080 to the control group and 480 to each treatment group at Session 1.

Table 3. Study Power by Sample Allocation at Session 2

Sample Allocation at Session 2		**Within-person Correlation**	**FDR**					**Unadjusted Power**
Control	**Treatment**		**0.05**	**0.1**	**0.15**	**0.2**	**0.25**	
287	287	0	0.60	0.71	0.83	0.89	0.91	0.68
880	250	0	0.80	0.89	0.94	0.95	0.97	0.85
1,200	230	0	0.77	0.87	0.90	0.94	0.95	0.81
1,520	210	0	0.88	0.92	0.94	0.96	0.98	0.88
1,840	190	0	0.72	0.84	0.88	0.92	0.93	0.78
287	287	0.2	0.64	0.77	0.83	0.89	0.90	0.73
880	250	0.2	0.95	0.95	0.97	0.99	0.99	0.96
1,200	230	0.2	0.89	0.94	0.95	0.98	1.00	0.91

1,520	210	0.2	0.87	0.92	0.94	0.96	0.98	0.89
1,840	190	0.2	0.85	0.88	0.94	0.94	0.99	0.87
287	287	0.4	0.85	0.91	0.96	0.98	0.99	0.89
880	250	0.4	0.97	0.98	0.99	0.99	1.00	0.98
1,200	230	0.4	0.97	1.00	1.00	1.00	1.00	0.97
1,520	210	0.4	0.95	0.97	0.97	0.97	0.98	0.95
1,840	190	0.4	0.91	0.94	0.97	0.97	0.97	0.91
287	287	0.6	0.97	0.98	0.98	1.00	1.00	0.98
880	250	0.6	1.00	1.00	1.00	1.00	1.00	1.00
1,200	230	0.6	1.00	1.00	1.00	1.00	1.00	1.00
1,520	210	0.6	1.00	1.00	1.00	1.00	1.00	1.00
1,840	190	0.6	1.00	1.00	1.00	1.00	1.00	1.00
287	287	0.8	1.00	1.00	1.00	1.00	1.00	1.00
880	250	0.8	1.00	1.00	1.00	1.00	1.00	1.00
1,200	230	0.8	1.00	1.00	1.00	1.00	1.00	1.00
1,520	210	0.8	1.00	1.00	1.00	1.00	1.00	1.00
1,840	190	0.8	1.00	1.00	1.00	1.00	1.00	1.00

Note: FDR = False Discovery Rate

4. Analysis

In the sections that follow, we describe our approach for three phases of analysis: In Phase 1, we will examine how reactions to warnings vary between GHW and SG warnings. In Phase 2, we will conduct a longitudinal analysis to examine the extent to which changes in beliefs vary between those exposed to GHW vs. those exposed to SG warnings. Finally, in Phase 3 we will assess variation in recall of warnings between those exposed to GHW vs. SG warnings. Note that the level of detail provided for each analysis Phase is commensurate with the level of complexity of the according analyses. The Phase 2 analyses involve longitudinal assessments of belief change and include methodological considerations around variable measurement and scaling that warrant more detailed explanation than Phase 1 and Phase 3 analyses.

4.1. Phase 1: Reactions to Warnings (Post-Exposure)

For the Phase 1 analysis, we will conduct comparisons of means and proportions for key reaction measures related to the warnings. Note that participants in the control condition will be exposed to a random selection of one of four Surgeon General's (SG) warnings; thus, each analysis will compare reaction measure means or proportions for a particular treatment condition to the means or proportions of the control group as averaged across the four SG warnings (i.e., we will compare treatment scores to a single control group score, rather than conducting separate analyses for each SG warning within the control condition). Table 4 provides a summary of the dependent variables, variable treatment, hypothesis, and analysis approach for each variable. Each analysis in this Phase will be repeated for each treatment-control comparison, for a total of

16 analyses per dependent variable.

Table 4. Phase 1 Analyses

Item #	Dependent Variable	Construct	Variable Treatment	Hypothesis	Analysis
B1	Before today, had you heard about the specific smoking-related health effect described in the warning? [Yes / No / I’m not sure]	New information	Dichotomous [Yes (0) vs. No / I’m not sure (1)]	H_0: proportion (%) responding that the warning provides new information (had not heard of the information contained in the warning prior to the experimental exposure) for those in the treatment condition = proportion (%) responding that warning provides new information for those in the control condition H_a: proportion (%) responding that warning provides new information for those in the treatment condition ≠ proportion (%) responding that warning provides new information for those in the control condition	Logistic regression
B12	To what extent did you learn something new from this warning that you did not know before? [7-pt scale from 1 (Not at all) to 7 (Very Much)]	Knowledge gain	Continuous	H_0: the mean response to B12 for those in the treatment group = the mean response to B12 for those in the control group. H_a: the mean response to B12 for those in the treatment group ≠ the mean response to B12 for those in the control group.	Linear regression
B10	How much does this warning make you	Thinking about Risks	Dichotomous [Somewhat / A	H_0: proportion (%) responding that the warning made them think about the health risks of	Logistic regression

Item #	Dependent Variable	Construct	Variable Treatment	Hypothesis	Analysis
	think about the health risks of smoking? [Not at all / A little / Somewhat / A lot]		lot (1) vs. Not at all / A little (0)]	smoking somewhat or a lot for those in the treatment condition = proportion (%) responding that warning made them think about the health risks of smoking somewhat or a lot for those in the control condition H_a: proportion (%) responding that the warning made them think about the health risks of smoking somewhat or a lot for those in the treatment condition ≠ proportion (%) responding that warning made them think about the health risks of smoking somewhat or a lot for those in the control condition	
B8_1	This statement is…[7-pt. scale from 1 (Not at all informative) to 7 (Very informative)] (B8_1)	Perceived informativeness	Continuous	H_0: the mean response to B8_1 for those in the treatment group = the mean response to B8_1 for those in the control group. H_a: the mean response to B8_1 for those in the treatment group ≠ the mean response to B8_1 for those in the control group.	Linear regression
B8_2	This statement is…[7-pt. scale from 1 (Hard to understand) to 7	Perceived understandability	Continuous	H_0: the mean response to B8_2 for those in the treatment group = the mean response to B8_2 for those in the control group.	Linear regression

Item #	Dependent Variable	Construct	Variable Treatment	Hypothesis	Analysis
	(Easy to understand)] (B8_2)			H_a: the mean response to B8_2 for those in the treatment group ≠ the mean response to B8_2 for those in the control group.	
B9	Would you say that this warning statement is an opinion or a fact? [Opinion / Fact]	Perceived factualness	Dichotomous [Fact (1) / Opinion (0)]	H_0: proportion (%) responding that the warning is a fact for those in the treatment condition = proportion (%) responding that warning is a fact for those in the control condition H_a: proportion (%) responding that the warning is a fact for those in the treatment condition ≠ proportion (%) responding that warning is a fact for those in the control condition	Logistic regression

To test the hypotheses, for each outcome, we will estimate a regression model of the following general form:

$$\text{Outcome} = f(\text{Condition, Age, Smoking Status})$$

where Outcome is a measure of reaction to the warning, Condition is a dichotomous indicator for a treatment versus control condition, Age is a categorical variable for age group (i.e. youth aged 13-17; young adults aged 18-24; and adults aged 25+), and Smoking Status is an indicator for current smoking versus not current smoking (in the adult and young adult samples non-current smokers are never smokers and in the youth sample non-current smokers are those youth susceptible to smoking). These models will include covariates for age and smoking status group to account for potential associations between age, smoking status and outcomes of interest.

The coefficient from the Condition variable indicates whether the outcome was significantly higher among those exposed to a GHW as compared to those exposed to a SG warning. This general model will be repeated for each of 16 treatment versus control group comparisons. All regressions, both logistic and linear will be estimated in Stata version 14.1 and will be estimated using Stata's robust standard errors.

As a supplement to the analyses above, we will conduct parallel analyses stratified by age group (i.e. youth aged 13-17; young adults aged 18-24; and adults aged 25+) and smoking status (current smoker versus non-smoker) to examine potential effects within each age and smoking status group. Of note is that this study is not sufficiently powered to detect within-age-group or smoking status differences, and so results from the stratified analyses should be interpreted with caution (i.e. a non-significant finding within an age or smoking status group may reflect lack of statistical power).

A total of 96 statistical tests will be conducted in Phase 1 (not including supplemental age-stratified analyses). To account for the possibility of falsely detecting a significant result (i.e. Type 1 error) arising from multiple statistical tests, we will control for the False Discovery Rate (FDR) using the Benjamini-Hochberg procedure (assuming a two-tailed test and FDR of 0.05)[1]. The Benjamini-Hochberg procedure involves ranking all the p-values from a family of tests from smallest to largest. The smallest p-value has a rank of i=1, the next smallest has i=2, etc. The next step is comparing each individual p-value to its Benjamini-Hochberg critical value, $(i/m)Q$, where i is the rank, m is the total number of tests, and Q is the FDR you choose. The largest p-value that has $P<(i/m)Q$ is statistically significant, and all of the p-values smaller than it are also statistically significant, even the ones that are not less than their Benjamini-Hochberg critical value. In other words, once a p-value in the list satisfies $P\geq(i/m)Q$, then no other p-values of that value or larger are considered statistically significant (and all less than that value are statistically significant).

There is little guidance on the best FDR to use in a study. Note that for an FDR of 0.05, the smallest p-value needs to be less than what would be the conservative Bonferonni correction (0.05/m), i.e., when i=1, then the Benjamini-Hochberg critical value is (1/m)*0.05. At an FDR of 0.05, the Benjamini-Hochberg critical value becomes slightly less conservative than a Bonferonni cut-off if p-values are less than this cut-off. However, if no p-values are less than 0.05/m, then no results are statistically significant. Thus, an FDR of 0.05 is conservative, like a Bonferonni correction. Accordingly, we will report results indicating statistical significance using an FDR of 0.05 (most conservative) and using no adjustment for multiple comparisons (least conservative).

We do not anticipate substantial item non-response, nor do we expect that patterns of item non-response would vary significantly among study conditions. Thus, we will plan to use pairwise deletion to include all available data for each particular analysis. We will examine the data for issues of item non-response and differential item non-response and adjust our approach for handling missing data accordingly.

4.2. Phase 2 Analysis: Condition-level Comparisons of Change in Beliefs

Model 1: Change in Beliefs from Session 1 to Session 2

For the Phase 2 analysis, we will conduct treatment vs. control comparisons of change in beliefs about the negative health consequences of smoking contained in the warnings. Note that participants in the control condition will be exposed to a random selection of one of four Surgeon General's (SG) warnings; thus, each analysis will compare the change in belief scores between a particular treatment condition and the control group as averaged across the four SG warnings (i.e., we will compare treatment scores to a single control group score, rather than conducting separate analyses for each SG warning within the control condition).

For each experimental condition, the survey includes an item or series of items in which respondents are asked to rate their level of agreement with a statement about a negative health consequence corresponding to the warning for that condition. The number of items associated with a particular warning ranges from 1 to 4. These items are assessed once during Session 1 before stimuli exposure, and then again following second stimuli exposure in Session 2. Table 5 provides a summary of the comparisons, dependent variables, and analysis approach for each of the Phase 2 analyses.

Table 5. Phase 2 Analyses

Analysis #	Comparison	Dependent Variable(s) [All 5-level “Strongly disagree” to “Strongly agree” response options]	Analysis

1	Condition 1 vs. Control (0)	A1_1. Cigarettes are addictive	Ordinal logistic regression
2	Condition 2 vs. Control (0)	A2_1. Tobacco smoke can harm your children	Ordinal logistic regression
3	Condition 3 vs. Control (0)	A3_1. Smoking can kill you	Ordinal logistic regression
4	Condition 4 vs. Control (0)	A4_1. Smoking causes fatal lung disease in nonsmokers	Ordinal logistic regression
5	Condition 5 vs. Control (0)	A5_1. Quitting smoking now greatly reduces serious risks to your health	Ordinal logistic regression
6	Condition 6 vs. Control (0)	A6_1. Smoking causes head cancer A6_2. Smoking causes neck cancer	Linear regression (if scaled); ordinal logistic regression (if items treated separately)
7	Condition 7 vs. Control (0)	A7_1. Smoking causes bladder cancer, which can lead to bloody urine A7_2. Smoking causes bladder cancer	Linear regression (if scaled); ordinal logistic regression (if items treated separately)
8	Condition 8 vs. Control (0)	A8_1. Smoking during pregnancy stunts fetal growth	Ordinal logistic regression
9	Condition 9 vs. Control (0)	A9_1. Smoking causes heart disease A9_2. Smoking causes strokes A9_3. Smoking clogs arteries A9_4. Smoking clogs arteries, which causes heart disease A9_5. Smoking clogs arteries, which causes strokes	Linear regression (if scaled); ordinal logistic regression (if items treated separately)
10	Condition 10 vs. Control (0)	A10_1. Smoking causes COPD, a lung disease that can be fatal A10_2. Smoking causes COPD A10_3. Smoking causes a lung disease that can be fatal	Linear regression (if scaled); ordinal logistic regression (if items treated separately)
11	Condition 11 vs. Control (0)	A10_1. Smoking causes COPD, a lung disease that can be fatal A10_2. Smoking causes COPD	Linear regression (if scaled); ordinal logistic

		A10_3. Smoking causes a lung disease that can be fatal	regression (if items treated separately)
12	Condition 12 vs. Control (0)	A12_1. Smoking reduces blood flow, which can cause erectile dysfunction A12_2. Smoking reduces blood flow A12_3. Smoking can cause erectile dysfunction	Linear regression (if scaled); ordinal logistic regression (if items treated separately)
13	Condition 13 vs. Control (0)	A13_1. Smoking reduces blood flow to the limbs, which can require amputation A13_2. Smoking reduces blood flow to the limbs A13_3. Smoking can lead to amputation	Linear regression (if scaled); ordinal logistic regression (if items treated separately)
14	Condition 14 vs. Control (0)	A14_1. Smoking causes type 2 diabetes, which raises blood sugar. A14_3. Smoking can cause Type 2 Diabetes	Linear regression (if scaled); ordinal logistic regression (if items treated separately)
15	Condition 15 vs. Control (0)	A15_1. Smoking causes age-related macular degeneration, which can lead to blindness A15_2. Smoking causes age-related macular degeneration A15_3. Smoking can lead to blindness	Linear regression (if scaled); ordinal logistic regression (if items treated separately)
16	Condition 16 vs. Control (0)	A16_1. Smoking causes cataracts, which can lead to blindness A16_2. Smoking causes cataracts	Linear regression (if scaled); ordinal logistic regression (if items treated separately)

The items being used to measure beliefs have Likert-type response scales. Conceptually, the response categories for a Likert response scale represent an underlying belief continuum. For warning statements with multiple corresponding items, we will assess whether to scale the items, using the following protocol:

1) Run a test of internal consistency reliability using Cronbach's alpha (Cronbach, 1951) on all of the items in a domain. If the test indicates "modest" reliability of alpha >= 0.70 (Nunnally & Bernstein, 1994), scale the items.

2) If alpha < 0.70, but all item-total correlations (i.e. the correlation between the item score and the overall scale score) are >= 0.4, scale the items (Item-total correlations of between 0.30—0.40 and greater have been suggested as sufficiently discriminating (Nunnally & Bernstein, 1994; Traub, 1994; Leong & Austin, eds., 2006)
3) If criteria 1 and 2 are not met, determine whether the scale alpha would increase to >= 0.70 if any items were deleted from the scale (i.e. using Stata's "alpha" command with "item" option specified). If the alpha value threshold would be met by dropping an item or items:
    a. Drop those items to form a scale with alpha >=0.70
    b. Also run analyses of each item individually
4) Otherwise, run analyses of each item individually

We are measuring beliefs at each study session. The above procedure to inform the scaling of the belief items will be made on the beliefs assessed at Session 1 (baseline) and then applied to the beliefs measured at the remaining sessions. We do not expect the measurement structure of beliefs to change over a longitudinal sample over a relatively short period of time. Assuming attrition across study sessions, basing the measurement structure on the Session 1 sample utilizes the largest available sample.

To determine the GHW's immediate impact on targeted beliefs, we will examine the extent to which pre-post differences in beliefs vary between those exposed to GHW (treatment) and SG warnings (control). The general form of this analysis approach is as follows:

$$(\text{Belief}^{T}_{S2} - \text{Belief}^{T}_{S1}) - (\text{Belief}^{C}_{S2} - \text{Belief}^{C}_{S1})$$

where Belief represents the average value (for continuous variables) or probability of being in a higher response category (for ordinally-treated variables), among those in a Treatment (*T*) or Control (*C*) group, at Session 2 (*S2*) or Session 1 (*S1*).

For each treatment versus control comparison, will test hypotheses of the following general form:

- H_0: Average pre-post difference in belief score for those in the treatment condition = average pre-post difference in belief score in the control condition
- H_a: Average pre-post difference in belief score for those in the treatment condition ≠ average pre-post difference in belief score in the control condition

To test the hypotheses for Phase 2 analyses, for each outcome we will estimate a regression model of the following general form:

$$\text{Belief} = f(\text{Condition, Session, Condition*Session, Age, Smoking Status})$$

where Belief is a measure of agreement with a statement (or set of statements) about the health effects of cigarette smoking, Condition is a dichotomous indicator for a treatment versus control condition, Session is a dichotomous indicator for Session 2 versus Session 1, Condition*Session is the interaction between study condition and study session, Age is a categorical variable for age group (i.e. youth aged 13-17; young adults aged 18-24; and adults aged 25+), and Smoking Status is an indicator for current smoking versus not current smoking (in the adult and young adult samples non-current smokers are never smokers and in the youth sample non-current smokers are those youth susceptible to smoking). These models will include covariates for age and smoking status group to account for potential associations between age, smoking status and outcomes of interest. The key variable of interest in these models is the interaction term, Condition*Session. The coefficient on Condition*Session indicates whether the pre-post change in belief was greater among respondents exposed to a GHW as compared to those exposed to a SG warning. This general model will be repeated for each of 16 treatment versus control group comparisons.

For those statements which have multiple corresponding belief items that can be scaled into a single continuous variable, we will test these hypotheses using linear regression. For statements with single ordinal Likert-type belief items or for which multiple items are not scalable, we are testing hypotheses of the form that treatment (being exposed to GHW) is associated with a greater pre-post change in level of the ordinal dependent variable than being in the control group (being exposed to an SG warning). Thus, for these items we will use ordinal logistic regression. All regressions, both logistic and linear will be estimated in Stata version 14.1 and will be estimated using Stata's robust standard errors.

In the cases where the dependent variable is continuous and a linear regression model is estimated, the interaction term (Session*Condition) represents the difference in difference of the means or treatment effect. However, in a non-linear model, such as when the dependent variable is ordered and we estimate an ordinal regression model, the coefficient of the interaction term is not a direct measure of the treatment effect due to the non-linear model. As noted in Puhani (2012), in a non-linear model with a strictly monotonic transformation function of a linear index, the sign of the coefficient of the interaction term is equal to the sign of the treatment effect. Testing the significance of the interaction term in the non-linear model is best done via bootstrapping (Puhani, 2012).

As a supplement to the analyses above, we will conduct parallel analyses stratified by age group (i.e. youth aged 13-17; young adults aged 18-24; and adults aged 25+) and smoking status (current smoker versus non-smoker), to examine potential effects within each age and smoking status group. Of note is that this study is not sufficiently powered to detect within-age or smoking status group differences, and so results from the stratified analyses should be interpreted with caution (i.e., a non-significant finding within an age or smoking status group may reflect lack of statistical power).

A total of 16 statistical tests will be conducted in Phase 2 (not including supplemental age-stratified analyses). To account for the possibility of falsely detecting a significant result (i.e. Type 1 error) arising from multiple statistical tests, we will control for the False Discovery Rate (FDR) using the Benjamini-Hochberg procedure (assuming a two-tailed test and FDR of 0.05).

We expect some level of overall attrition between Session 1 and Sessions 2 and 3. Although unlikely given the nature of the experimental procedure, there is a potential that the rate of attrition could vary between treatment and control groups, resulting in biased estimates of the effect of the GHWs. To assess potential problems resulting from differential attrition, we will calculate and report rates of overall attrition (i.e., the proportion of Session 1 participants randomly assigned to a treatment or control group for whom Session 2 / Session 3 data are not available) and differential attrition (i.e., the difference in attrition rates between treatment and control groups). We will report overall and differential rates of attrition for each of 16 treatment groups, assessed at Session 2 and Session 3. We have no *a priori* threshold for determining an acceptable level of attrition bias. Nevertheless, to contextualize findings, we can compare attrition rates to benchmarks for randomized controlled trials established by the Department of Education's What Works Clearinghouse (Deke et al., 2015).

We do not anticipate substantial item non-response, nor do we expect that patterns of item non-response would vary significantly among study conditions. Thus, we will plan to use pairwise deletion to include all available data for each particular analysis. We will examine the data for issues of item non-response and differential item non-response and adjust our approach for handling missing data accordingly.

Model 2: Change in Beliefs at Session 3

To determine the GHW's sustained impact on targeted beliefs, we will conduct parallel analyses to the Model 1 analyses described above, but with a Session indicator variable that indicates Session 3 versus Session 1 (as opposed to Session 2 vs. Session 1). The general form of this analysis approach is as follows:

$$(\text{Belief}^{T}_{S3} - \text{Belief}^{T}_{S1}) - (\text{Belief}^{C}_{S3} - \text{Belief}^{C}_{S1})$$

where belief represents the average value (for continuous variables) or probability of being in a higher response category (for ordinally-treated variables), among those in a Treatment (*T*) or Control (*C*) group, at Session 3 (*S3*) or Session 1 (*S1*).

The specific functional form of these models, hypothesis tests, and interpretation of coefficients are identical to those described for Model 1, with the exception that with Model 2 we are examining differences in Session 3 versus Session 1 belief values.

4.3. Phase 3: Warning Label Recall

For the Phase 3 analysis, we will conduct comparisons of the proportion of respondents accurately recalling (at Session 3) the warning that they were exposed to at Sessions 1 and 2. We assess recall with the following item:

E1. You recently took a survey in which you were shown a cigarette pack and advertisement with a warning on it. Which label do you remember seeing?

1. LABEL 1
2. LABEL 2
3. LABEL 3
4. LABEL 4
5. None of these
6. I don't remember

Respondents in the control condition are shown each of the 4 SG labels; those in each treatment condition are shown the GHW that they were exposed to earlier along with 3 randomly selected additional GHWs. Thus, each respondent is shown one warning that they were exposed to earlier in the study, and 3 warnings (of the same type—GHW or SG) that they had not been exposed to. We will construct an indicator variable such that 1 = accurate recall of the warning to which the respondent was exposed and 0 = inaccurate or lack of recall (i.e., false recall of any of the 3 warnings not shown earlier in the study or a response of "None of these" or "I don't remember").

For the Phase 3 analysis, we will test hypotheses of the following general form:

- H_0: proportion (%) of those in the treatment condition accurately recalling the warning = proportion (%) of those in the control condition accurately recalling the warning

- H_a: proportion (%) of those in the treatment condition accurately recalling the warning ≠ proportion (%) of those in the control condition accurately recalling the warning

Since the recall measure is dichotomous, we will test this hypothesis using logistic regression of the following form:

$$\text{Recall} = f(\text{Condition, Age, Smoking Status})$$

where Recall is a measure of accurate recall of the warning, Condition is a dichotomous indicator for a treatment versus control condition, Age is a categorical variable for age group (i.e. youth aged 13-17; young adults aged 18-24; and adults aged 25+) and Smoking Status is an indicator for current smoking versus not current smoking (in the adult and young adult samples non-current smokers are never smokers and in the youth sample non-current smokers are those youth susceptible to smoking). These models will include covariates for age and smoking status group to account for potential associations between age, smoking status and outcomes of interest. The coefficient from the Condition variable indicates whether accurate warning recall was significantly greater among those exposed to a GHW as compared to those exposed to a SG warning. This general model will be repeated for each of 16 treatment versus control group comparisons. All regressions will be estimated in Stata version 14.1 and will be estimated using Stata's robust standard errors.

As a supplement to the analyses above, we will conduct parallel analyses stratified by age and smoking status group, to examine potential effects within each age and smoking status group. Of note is that this study is not sufficiently powered to detect within-age or smoking status-group differences, and so results from the stratified analyses should be interpreted with caution (i.e. a non-significant finding within an age or smoking status group may reflect lack of statistical power).

A total of 16 statistical tests will be conducted in Phase 3 (not including supplemental age-stratified analyses). To account for the possibility of falsely detecting a significant result (i.e. Type 1 error) arising from multiple statistical tests, we will control for the False Discovery Rate (FDR) using the Benjamini-Hochberg procedure (assuming a two-tailed test and FDR of 0.05) (Benjamini & Hochberg, 1995).

We do not anticipate substantial item non-response, nor do we expect that patterns of item non-response would vary significantly among study conditions. Thus, we will plan to use pairwise deletion to include all available data for each particular analysis. We will examine the data for issues of item non-response and differential item non-response and adjust our approach for handling missing data accordingly.

References

Benjamini, Y. & Hochberg Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. J Roy Stat Soc. Series B, 57(1), pp. 289–300.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. psychometrika, 16(3), 297-334.

Deke, J., Sama-Miller, E., and Hershey, A. (2015). Addressing Attrition Bias in Randomized Controlled Trials: Considerations for Systematic Evidence Reviews." Washington, DC: Office of

Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, 2015.

Leong, F. T., & Austin, J. T. (2006). The psychology research handbook: A guide for graduate students and research assistants. Sage.

Nunnally, J. C., & Bernstein, I. H. (1994). Psychological theory. New York, NY: MacGraw-Hill.

Puhani, P. A. (2012). The treatment effect, the cross difference, and the interaction term in nonlinear "difference-in-differences" models. Economics Letters, 115(1), 85-87.

Traub, R. E. (1994). Reliability for the social sciences: Theory and applications (Vol. 3). SAGE publications.