Complete Title: An evaluation of social media warning labels for teens and young adults: A randomized controlled trial
Short Title: Reactions to social media warning labels
Protocol Date: August 1, 2025
NCT #: 07199660

Study Principal Investigator:

Anna H. Grummon
Stanford University
email: agrummon@stanford.edu

<u>Protocol Synopsis</u>

Study Title	Social Media Warning Labels
Funder	NA
Clinical Phase	NA
Study Rationale	• Among teens and young adults, average daily social media use is high and contributes to poor mental health and other adverse outcomes. Warning labels are a low-cost, highly scalable strategy for informing consumers about product harms and discouraging unhealthy behaviors. • It remains unknown how teens and young adults respond to social media warnings, including which warning topics may be most effective at discouraging social media use and raising awareness of the harms of social media.
Study Objective(s)	The primary objective is to evaluate whether different social media warnings are perceived as effective at discouraging social media use and raising awareness of the harms of social media. A secondary objective is to compare perceived message effectiveness of warnings refined using artificial intelligence (AI) vs. those not refined using AI.
Study Design	Within-subjects randomized experiment.
Subject Population key criteria for Inclusion and Exclusion:	Inclusion Criteria 1. 13 to 29 years old 2. Reside in the US 3. Can read and speak English
Number of Subjects	1,000
Study Duration	Each subject's participation will last approximately 10 minutes. The enrollment period is expected to last ~4 weeks.
Study Phases	There are two phases: (1) <u>Screening</u>: screening for eligibility and obtaining consent and (2) <u>Intervention</u>: study intervention/experimental treatment.
Efficacy Evaluations	The primary outcome is perceived message effectiveness for discouraging social media use. It is measured with 1 survey item. The secondary outcome is perceived awareness of the harms of social media use. It is measured with 1 survey item.
Statistical and Analytic Plan	Primary outcome • First, we will examine the effect of each social media warning topic on perceived message effectiveness compared to control. Second, we will examine the effect of each potential mandatory social media warning topic compared to a voluntary social media

	warning. Third, we will test whether the effects of the social media warning topics in this study with human participants differ from the effects observed in a separate experiment conducted with artificial intelligence (AI) personas. Fourth, we will compare social media warnings we refined using AI personas to those we selected *a priori* and did not refine using AI. Secondary outcomes • We will examine the effect of each social media warning topic on awareness of the harms of social media compared to the control topic. We will also examine the effect of each potential mandatory social media warning topic compared to a voluntary social media warning.
Data and Safety Monitoring Plan	• The principal investigators are responsible for data quality management and ongoing assessment of safety.

Introduction

The primary goal of the analyses described here is to use data we collected through an online randomized experiment to examine consumer responses to different social media warning messages. These analyses examine the effect of warning topic (e.g., depression and anxiety, body image, sleep) on perceived message effectiveness (primary outcome) and perceived awareness of the harms of social media (secondary outcome).

A secondary goal of the analyses described here is to test whether the effects of the social media warning topics in this study with human participants differ from the effects observed in a separate experiment conducted with artificial intelligence (AI) personas. Another secondary goal is to examine whether warnings refined using AI are perceived as more effective than those selected *a priori* and not refined using AI.

This analysis plan pre-specifies the analyses before collecting data and therefore serves as our ex-ante planned analysis.

Study Protocol

Participants will complete a within-subjects online randomized experiment. After providing informed consent, participants will view and rate messages on perceived message effectiveness (primary outcome) and perceived awareness of the harms of social media (secondary outcome). Participants will view messages about 9 topics: 8 warning topics and 1 control topic. The 8 warning topics include 7 potential mandatory social media warnings (including the topics of depression and anxiety, body image, addiction, sleep, mental health harms to children, not been proven safe, and California's proposed warning) and 1 voluntary warning (similar to a message used on TikTok to encourage users to take breaks). For each topic, participants will view 1-2 messages and respond to survey items about that message. All messages will be shown in random order.

Statistical Considerations

General Principles

We will use a two-sided critical alpha of 0.05 as a criterion for all tests of statistical significance. All confidence intervals presented will be 95% and two-sided. We will follow intention-to-treat principles, including all participants who fully enroll in the study. We expect minimal missing data based on prior similar studies[1–3] but missing data will be accounted for under maximum likelihood assumptions in our mixed effects regression analysis of primary and secondary outcomes.

Primary Outcome

The primary outcome is perceived message effectiveness for discouraging social media use. We will measure perceived message effectiveness with 1 item adapted from prior studies,[4,5] "How much does this message discourage you from wanting to use social media?" Response options will range from not at all (1) to a great deal (5).

Secondary Outcomes

The secondary outcome is reported increased awareness of the harms of social media. We will measure perceived awareness of the harms of social media with 1 item, "How much does this message increase your awareness of the harms of using social media?" Response options will range from not at all (1) to a great deal (5).

Statistical Methods

We plan to present results in two papers.

Paper 1 Analyses

1. Analyses of the primary outcome
   a. We will use mixed effects linear regression to **evaluate the effect of each warning topic compared to the control topic on the primary outcome of perceived message effectiveness.** We will regress perceived message effectiveness on a set of indicator variables representing each social media warning topic (e.g., depression and anxiety, negative body image, sleep, etc.), excluding the control topic as the referent. We will treat the intercept as random to account for repeated measures within participants. The coefficients on the warning topics will give the average difference in mean perceived message effectiveness between each warning topic and the control topic. Given the exploratory nature of the study, we do not plan to correct *p*-values for multiple comparisons.
   b. We will use the same mixed effects linear regression model to **evaluate the effect of each potential mandatory warning topic compared to the voluntary warning.**

   c. In exploratory analysis, we will test whether the effects of the warning topics on the primary outcome are possibly moderated by age (treated continuously), gender (male vs. female), and amount of social media use (treated continuously). To test for moderation, we will use mixed effects linear regression, regressing perceived message effectiveness on indicator variables representing each social media warning topic (excluding the control as the referent), the possible moderator, and the interactions between the possible moderator and the warning topics. We will center variables prior to analysis. We will use separate models for each moderator. We will test the joint significance of the interaction terms.
   d. We will also **descriptively rank the warning topics** on the primary outcome of perceived message effectiveness. We will estimate mean perceived message effectiveness for each social media warning topic (averaging across messages for each topic) and rank those means.
2. Analyses of the secondary outcome:
   a. We will use mixed effects linear regression to **evaluate the effect of each social media warning topic compared to the control topic on the secondary outcome of awareness of the harms or social media.** We will use the same approach as described for the primary outcome (see no. 1 above).
   b. We will use the same mixed effects linear regression model to **evaluate the effect of each potential mandatory warning topic compared to the voluntary warning.**
   c. We will **descriptively rank the social media warning *topics*** on the secondary outcome of perceived awareness of the harms of social media. We will estimate mean perceived awareness of the harms of social media for each topic (averaging across messages for each topic) and rank those means.

Paper 2 Analyses

1. Analyses of the primary outcome:
   a. We will test whether the effects of the social media warning topics in this study with human participants differ from the effects observed in a separate experiment conducted with artificial intelligence (AI) personas. We will compare the coefficient on each warning topic estimated in this study to the corresponding coefficient estimated in the separate experiment conducted with AI personas. We will pool data from this study with the data from the separate experiment conducted with AI personas and run a mixed effects linear regression, regressing perceived message effectiveness on indicator variables for each warning topic (excluding the control as the referent), an indicator variable for sample (i.e., whether the observation is from this study vs. the AI study), and the interaction between warning topic and sample. All variables will be centered. We will examine the coefficients on the interaction terms.

b. We will compare how our human participants rate social media warnings we refined using AI personas to those we selected *a priori* and did not refine using AI. We will make this comparison both overall (pooling across all warning topics) and within each warning topic. These analyses will only include ratings from the 6 warning topics for which we tested both AI-developed and human-developed messages (depression and anxiety, negative body image, addiction, sleep, mental health harms to children, and not been proven safe).

   i. First, we will used mixed effects regression, regressing perceived message effectiveness on an indicator variable for message source (AI-developed vs. human-developed) and treating the intercept as random. The coefficient on the indicator variable will indicate the average difference in means between AI-developed vs. human-developed messages.

   ii. Second, we will run a set of mixed effects regressions, one for each warning topic. Each regression will regress perceived message effectiveness on an indicator variable for message source (AI-developed vs. human-developed), treating the intercept as random. The coefficient on the indicator variable will indicate the average difference in means between AI-developed vs. human-developed messages.

Sample Size Needs

We plan to collect data from 1,000 participants. We used G*Power[6] to estimate sample size needs to detect an effect of each social media warning topic vs. control using ANOVA repeated measures (testing of within-subject factors). Assuming alpha=0.05, 2 measurements (given each contrast will compare 2 topics to one another), and a conservative correlation among repeated measures of 0.5 (based on prior studies[7–9]), a sample size of 900 would yield 85% power to detect effects of *f*=0.05 (*d*=0.10) or larger. This effect size would be considered small[10] and is conservative based on our prior message development studies.[8,9,11] We will recruit 1,000 participants to account for potential missing data.

Exclusions and Outliers

We will exclude human participants who do not complete the survey or who complete the survey implausibly quickly (defined as <1/3 of the median completion time). We will winsorize outlier values defined as more than 3 interquartile ranges below or above the 25th and 75th percentiles of the observed distributions (i.e., Tukey far outliers). These participants and winsorized observations will be included in the analysis.

References

1. Grummon AH, Gibson LA, Musicus AA, Stephens-Shields AJ, Hua SV, Roberto CA. Effects of 4 interpretive front-of-package labeling systems on hypothetical beverage and snack selections: A randomized clinical trial. *JAMA Network Open*. 2023;6(9):e2333515. doi:10.1001/jamanetworkopen.2023.33515

2. Taillie LS, Bercholz M, Prestemon CE, et al. Impact of taxes and warning labels on red meat purchases among US consumers: A randomized controlled trial. *PLOS Medicine*. 2023;20(9):e1004284. doi:10.1371/journal.pmed.1004284

3. Grummon AH, Musicus AA, Moran AJ, Salvia MG, Rimm EB. Consumer reactions to positive and negative front-of-package food labels. *Am J Prev Med*. 2023;64(1):86-95. doi:10.1016/j.amepre.2022.08.014

4. Baig SA, Noar SM, Gottfredson NC, Boynton MH, Ribisl KM, Brewer NT. UNC Perceived Message Effectiveness: Validation of a brief scale. *Ann Behav Med*. 2019;53(8):732-742. doi:10.1093/abm/kay080

5. Baig SA, Noar SM, Gottfredson NC, Lazard AJ, Ribisl KM, Brewer NT. Message perceptions and effects perceptions as proxies for behavioral impact in the context of anti-smoking messages. *Prev Med Rep*. 2021;23:101434. doi:10.1016/j.pmedr.2021.101434

6. Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*. 2007;39(2):175-191. doi:10.3758/BF03193146

7. Grummon AH, Hall M, Taillie L, Brewer N. How should sugar-sweetened beverage health warnings be designed? A randomized experiment. *Prev Med*. 2019;121:158-166.

8. Grummon AH, Lee CJY, D'Angelo Campos A, et al. Health harms that discourage alcohol consumption: A randomized experiment of warning messages. *Addictive Behaviors*. 2024;159:108135. doi:10.1016/j.addbeh.2024.108135

9. Grummon AH, Ruggles PR, Greenfield TK, Hall MG. Designing effective alcohol warnings: Consumer reactions to icons and health topics. *Am J Prev Med*. 2023;64(2):157-166. doi:10.1016/j.amepre.2022.09.006

10. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. Academic press; 2013.

11. Hall MG, Lazard AJ, Grummon AH, et al. Designing warnings for sugary drinks: A randomized experiment with Latino and non-Latino parents. *Prev Med*. 2021;148:106562. doi:10.1016/j.ypmed.2021.106562