

An Assessment of “Proposition 65 Clear and Reasonable Warning Regulations Study”

Alexander J. Oliver, MA
Adam B. Schaeffer, PhD

Summary

The American Chemistry Council engaged Evolving Strategies LLC to assess the “Proposition 65 Clear and Reasonable Warning Regulations Study” on the “effectiveness” of proposed labeling changes under Proposition 65 (hereafter referred to as the “study”). The study was conducted by the UC Davis Extension Collaboration Center at the request of the Office of Environmental Health Hazard Assessment (OEHHA) to determine “whether the existing or proposed warnings are *more helpful as a clear warning of chemical exposure*” (p.5). The study fails — on many fronts — to demonstrate that the proposed changes would be more effective at advancing the intent of the law.

There are three broad categories in which the research falls short on its own terms —

- the survey sample,
- the survey instrument design, and
- the survey execution.

First, the sampling procedure used by the researchers is ad-hoc and delivers a population that is irredeemably biased and unrepresentative. Rather than systematically surveying a representative sample of all California citizens, the researchers chose to intercept and survey only those residents who happened to be at one of 19 Department of Motor Vehicle (DMV) locations that massively overrepresent or underrepresent distant regions of the state. Further, the people available at a DMV differ substantially from the general population of California since residents can complete most essential DMV tasks online.

Based on the sampling procedure alone, the results cannot be taken at face value as an accurate reflection of public opinion in California.

Second, the research introduces substantial bias into the survey through the design and delivery of the survey instrument. The use of live interviews, questions that prime certain considerations, and biased outcome measures mean that we cannot use the results to draw valid conclusions.

The researchers act as authority figures in administering the survey and communicate to respondents many of the results that they expect to receive, which biases the respondents' answers. The researchers present statements of *opinion* as statements of *fact* with respect to Proposition 65, and they use language that moves respondents toward some (preferred, by implication) answer. And the researchers provide biased response options that render the results of some questions meaningless.

Third, the researchers ask respondents to self-predict the efficacy of labeling, despite extensive social science research that demonstrates self-response bias renders such self-reported measures ineffective in determining public service interventions.

The repetitiveness and length of the survey also calls into question the quality of the data. Under the substantial attentional and cognitive burdens imposed on them by the survey instrument, respondents may answer questions differently than they would otherwise – again biasing any inference about the effectiveness of the labeling and interventions.

There is also a serious methodological problem with the use of a survey, *per se*, to answer the question of whether the new labeling proposals are more effective. The researchers ask respondents to self-predict the efficacy of alternate labeling, despite extensive social science research that demonstrates such self-reported measures are not capable of demonstrating the effectiveness of public service interventions.

Researchers can only draw conclusions about the effectiveness of interventions based on results from a randomized-controlled trial testing alternative approaches (such a study is described in the final section). The study examined here uses introspection and self-reporting, which have proven inaccurate or even misleading across many different literatures.

Determining the effectiveness of public health interventions is a complicated and difficult problem. The standard and accepted practice is for researchers to randomly assign different treatments to a study's subjects and observe the outcomes of interest (which would include increased knowledge of and assessment of relative risks from various chemicals) in a randomized-controlled trial of labeling approaches. This requires a systematic sampling of research subjects and a more considered, rigorous research design and execution than used in this study.

Sampling Procedure

Page 5 of the study states that “the survey was designed to solicit California’s general public opinion of whether the existing or proposed warnings are *more helpful as a clear warning of chemical exposure*” (italics added) and that “OEHHA elected to survey a broad sample of the general public that represented the demographics of the State of California.” However, the researchers in fact solicit *specific* public opinion from a *narrow* sample of California’s population because:

1. The geographic areas where the researchers surveyed are not representative of all areas where California’s population lives.
2. The locations where the researchers surveyed in each of these geographies are not representative of the area’s population. Further, the researchers surveyed different areas / locations on rolling basis, on different days and in different weeks.

The study has sampling bias and as a result its sample is not representative (Kruskal and Mosteller, 1980).

Unrepresentative Areas

The study states that “Department of Motor Vehicle (DMV) locations were selected for survey locations [...] In total, data were collected from 19 DMV offices across urban and rural areas in the Central Valley, Southern California, San Francisco Bay Area, and Northern California” (p. 6). The researchers do not report how they selected these 19 DMV locations for survey locations.

But these surveyed areas are not representative of all areas where California's population lives. The number of respondents from each of California's statistical areas based on the 2010 census is not proportionate to its population size.

- The Los Angeles-Long Beach-Riverside, CA CSA (combined statistical area) comprises 48.0% of California's total population but just 20.1% of the survey sample— meaning the survey sample underrepresented this CSA by a factor of about 2.4. The researchers only surveyed at 3 DMVs here: Pasadena, Los Angeles, and Bell Gardens.
- The San Diego-Carlsbad-San Marcos, CA MSA (core base statistical area) comprises just 8.3% of California's total population but 21.2% of the survey sample — meaning the survey sample overrepresent this MSA by a factor of about 2.6. The researchers only surveyed 3 DMVs here: San Diego, San Diego Claremont, and El Cajon.
- The combination of the San Jose-San Francisco-Oakland, CA CSA, the Sacramento-Arden Arcade-Yuba City, CA CSA, the Modest, CA MSA, the Merced, CA MSA, the Chico, CA MSA, the Redding, CA MSA, the Red Bluff, CA MSA, and the Stockton, CA MSA comprises 31.9% of California's total population but 58.7% of the survey sample – meaning the survey sample overrepresented this CSA by a factor of about 1.8. The researchers surveyed 13 DMVs here: Manteca, Modesto, Turlock, Merced, Stockton, San Mateo, Redwood City, Santa Clara, San Jose, Oroville, Chico, Redding, and Red Bluff.

We hypothesize that the main reason for this 2.4-factor underrepresentation of the Los Angeles area and this 1.8-factor overrepresentation of the San Francisco-Sacramento area – northern California – is the researchers' *convenience*. The researchers could drive to any of the 13 DMVs that they chose in northern California in less than 2 hours from their base at Davis. The farthest two DMVs from Davis chosen in northern California are Merced to the south – 130 miles by road and an estimated 2 hours and 12 minutes' drive for the researchers – and Redding to the north – 154 miles by road and an estimated 2 hours and 18 minutes' drive for the researchers. The closest DMV to Davis chosen in northern California is Stockton to the southeast – just 61 miles by road and an estimated 57 minutes' drive for the researchers. But the researchers would need over 6 hours to drive from their base at Davis to any of the only 6 DMVs that they chose in southern California. The closet DMV to Davis chosen in southern California is Bell Gardens to the southeast – 407 miles by road and an estimated 6 hours and 7 minutes' drive for the researcher – or a 75 minute plane flight.

Unrepresentative Population

The study states that “The DMVs providing [sic] a ‘captured’ audience where people typically have time to complete a survey while waiting for their turn [...] Willing participants completed the survey as they waited” (p. 6). The researchers surveyed different DMVs in different areas on a rolling basis – on different days and in different weeks.

But this procedure does not indicate that the samples acquired at the 19 DMVs are representative of the population in that DMV's area. This representativeness could fail in at least three ways at a particular DMV.

1. People visiting a DMV may not be representative of the population of that DMV's area. According to the California DMV's website, the main “services provided by DMV offices include Vehicle Registration, Driver License and Identification (ID) Card

- Processing.” The California DMV *requires* people to visit a DMV office the *first* time they use their main services – vehicle registration, driver licenses, and ID cards – but offers *renewals* involving these main services both by mail *and* online. People who visit a DMV are either using one of their main services the first time or are people using one of them the second time but prefer to visit a DMV over mail or online, both of whom are unlikely to be representative of the population of a DMV’s area.
2. People who are “willing” to complete “the surveys as they waited” (p. 6) may not be representative of the population of that DMV’s area. In survey research, this is known as self-selection bias. Those at the DMV who are willing to complete the surveys as they waited may be different from those who refused on a combination of characteristics that make them different – and thus unrepresentative – of the population of that DMV’s area.
 3. People visiting a DMV on the day the researchers surveyed there may not be representative of the population of that DMV’s area. The researchers surveyed at Manteca, Modest, Turlock, and Merced on August 6-7; at San Mateo and Redwood City on August 12; at Santa Clara and San Jose on August 13; at Pasadena on August 17; at Los Angeles and Bell Gardens on August 18; at San Diego on August 19; at El Cajon on August 20; at Oroville and Chico on August 24; and at Redding and Red Bluff on August 25. The researchers do not report when they surveyed at Stockton. People who visit a DMV on different days of the week – Monday, Tuesday, Wednesday, Thursday, or Friday – or in different weeks of the month – the first, second, third, fourth, etc. – may be different from each other – and thus unrepresentative – of the population of that DMV’s area.

The researchers only address the problem of unrepresentativeness by showing that the demographics of the sample are similar to the census – and thus representative – on gender and race/ethnicity (page 7 of the report). Although the researchers collect data on the sample’s age, pregnancy status, and language, they do not compare these to the census.

The demographic comparisons that were made do not mitigate any of the three concerns about representativeness and bias in the survey results, as there are many important demographic and political characteristics not considered – including but not limited to education, income, general occupation, location of residence, marital status, parental status, etc. – let alone controlled for in the final survey estimates.

Survey Instrument

Live Researchers

The study’s researchers asked the survey instrument’s questions and recorded answers *live* instead of having *respondents* read the survey instrument’s questions and answer by themselves on paper or electronic device. The presence of live researchers interacting with respondents reduces a survey instrument’s validity – how likely it is that the survey instrument measures what the researchers intend – in two ways.

1. Live researchers induce a *Hawthorne effect*. A Hawthorne effect occurs when respondents know that researchers are observing their answers during a survey and change their answers to questions as a result (Mayo, 1948; Landsberger, 1958).

- Respondents can change their survey question answers in response to researcher observation for several reasons: respondents may react negatively or positively to the social interaction with the researchers; they may want to help the researchers by giving answers that support the researchers' hypotheses; or they may want to make the researchers perceive them as more socially desirable by giving answers that they think researchers view more favorably than others – a phenomenon called social desirability bias (Parsons, 1974; Steele-Johnson et. al., 2000; Crowne and Marlowe, 1960).
2. The study states that “Jodie Monaghan led the survey team. The student survey team members were Leigh Hiura, Rebecca Belloso, and Yadira Chavez” (p. 6) and a question on the survey instrument itself indicates “Researcher: Jodi, Yadira, Rebecca, Leigh, Other_____” (p. 64). Although there are four different members on the survey team asking questions and recording answers live, the study does not indicate that there is any protocol to make survey administration consistent among them. Survey administration is likely inconsistent across the four different survey team members as a result – especially since only “two members of the survey team were fluent Spanish speakers, able to engage Spanish-speakers” (p. 1) – which opens up the possibility that respondents' answers to questions may be influenced by which member of the survey team interviewed them.

A related problem may be the environment in which the researchers administered the survey. The study does not report *where inside* the DMV the researchers collected responses. Was it in a main area in the presence of other DMV patrons or in a separate area private from other patrons? Either of these locations may bias the answers of respondents, and without knowing which one the researchers used, it is impossible to assess just how much bias affected the outcome.

Priming

The study's survey instrument *primes* respondents by including certain questions. Priming occurs when earlier content in a survey – statements, questions, etc. – systematically influence respondents' answers to later questions (for the psychological basis of priming, see Meyer and Schvaneveldt, 1971). Priming reduces a survey instrument's validity – how likely it is that the survey instrument measures what the researchers intend. Priming enters this study's survey instrument in at least three ways.

1. Survey question 5 asks “How are you feeling today? Very negative, negative, neither negative nor positive, positive, very positive” (p. 64) and it appears before any of the study's questions of interest about the helpfulness of the signs, etc. and likely primes respondents. This question likely induces respondents to think about how they feel at the time asked by a member of the survey team and makes them self-aware and/or augments self-awareness of their current emotional state. This greater self-awareness of their emotional state – whether negative or positive – likely changes respondents' answers to the later questions of interest about the helpfulness of signs, etc. from how they otherwise would respond.
2. Survey question 6 asks “Proposition 65, the Safe Drinking Water and Toxic Enforcement Act of 1986, requires businesses to notify Californians about significant amounts of chemicals they may be exposed to. This enables the public to make informed decisions about protecting themselves from exposure to chemicals. Have you heard about Proposition 65 before today? Yes, no” (p. 65). This question

- appears before any of the study's questions of interest regarding the helpfulness of the signs, etc. and likely primes respondents. This question likely induces respondents to think not only about Proposition 65's intent – a factual statement (“requires businesses to notify Californians about significant amounts of chemicals they may be exposed to”) – but also about *justifications* for its existence – a *normative* statement (“this enables the public to make informed decisions about protecting themselves from exposure to chemicals”). This state of thought likely changes respondents' answers to the later questions of interest about the helpfulness of signs, etc., from how they otherwise would respond.
3. Survey question 7 asks, “Not all Prop 65 signs look the same, but they all provide warnings about dangerous chemicals. Here is an example of a Prop 65 warning sign. How often have you seen a sign like this before today? Several times a week, several times a month, a few times a year, never” (p. 64). The question appears before any of the study's questions of interest about the helpfulness of the signs, etc. and likely primes respondents. This question likely induces respondents to think negatively about “dangerous chemicals” and positively about Prop 65 signs that warn of those “dangerous chemicals.” This state of thought – with both negative and positive elements that contrast and conflict – likely makes respondents pay more attention to the survey than they would otherwise, which likely changes their answers to the later questions of interest about the helpfulness of signs, etc. from how they would otherwise respond.

Biased Questions (Measures) of Interest

As aforementioned, page 5 of the study states that “the survey was designed to solicit Californians' general public opinion of whether the existing or proposed warnings are more helpful as a clear warning of chemical exposure.” However, the survey *solicits* Californians about whether the existing or the proposed warnings are more helpful without alarming the public but cannot actually *determine* whether the existing or proposed warning labels are *in fact* more helpful while avoiding alarm because:

1. Three of the questions of interest about the helpfulness of signs, etc. suffer from *wording bias*.
2. One of the questions of interest regarding the meaning of the signs and impact on the respondent's mental state suffers from *response bias*.
3. Four of the questions of interest about helpfulness of signs, etc. suffer from *self-report bias*.

Wording Bias

Two sets of questions on the survey suffer from consistent *wording bias*. Questions 16 through 22 ask, “One sign includes the chemical names and the other sign refers generally to chemicals. Which sign is more helpful? Select one below” (p. 67) and questions 23 through 29 ask, “These two signs are identical in content, but arranged differently. Which sign is easier to read?” (p. 68).

In these two questions the researchers tell respondents what they want to measure and then measure it within the questions. Telling respondents what they want to measure changes how respondents will answer the researchers' questions – a form of “within-question” priming. This wording bias reduces the survey instrument's validity – how likely it is that the survey instrument measures what the researchers intend.

- In questions 16 through 22, researchers first state, “one sign includes the chemical names and the other sign refers generally to chemicals” before asking “which sign is more helpful?” As a result respondents evaluate the relative helpfulness of the two signs based on their specificity/generality of the description of chemicals and not based on the criteria they would use to do so naturally and this changes which of the two signs the respondents choose as more helpful.
- In questions 23 through 29, researchers first state, “these two signs are identical in content, but arranged differently” before asking, “which sign is easier to read?” As a result respondents evaluate the relative easiness of the two signs based on the arrangement of their content and not based on criteria they would use to do so naturally and this changes which of the two signs the respondents choose as easier to read.

One additional question has substantial wording bias: “If you wanted additional information, how likely are you to visit the website listed below?” (p.71).

First, the question does not ask the respondent simply “how likely are you to visit the website listed below?” The question instead begins by stating a hypothetical, “if you wanted additional information,” which suggests to the respondent that seeking more information is a socially desirable action here and implying that the website is the appropriate way to seek such information.

Research has demonstrated that individuals exaggerate the likelihood of engaging in social-desirable behaviors, and it is standard-practice to avoid indicating which response is socially desirable within a question (for a canonical reference, see Crowne and Marlowe, 1960).

Although researchers have provided respondents with a biased question containing a hypothetical indicating the socially-desirable answer, about 40 percent of respondents still report they would be unlikely to visit the website. Regardless, no survey question that measures self-reported intention can accurately predict real-world behavior such as visiting a website (see the last section of this document for more on self-reported intention versus actual behavior).

Response Option Bias

A key question the study asks is, “what does the triangular yellow/B&W symbol mean to you?” However, the response options provided for this question – combined with the implicit definition of “alarm” – bias this measure toward finding that no alarm is caused by the proposed symbol. It is likely that a less biased measure would result in a substantially different conclusion.

The question, as asked by the researchers, is irredeemably biased and cannot be used to draw any valid inferences about respondents' opinions of the symbol (Friedman et. al., 1981).

On page 36, the study explicitly states that the researchers intend this question “to assess whether inclusion of the triangular symbol creates alarm.” There are seven closed-ended response options:

1. Warning
2. Danger
3. Caution
4. Nothing
5. It confuses me
6. It scares me

7. It gets my attention
8. Other (please specify) _____

Responses 1-4 are logical answers appropriate to the question; the researchers take these responses to mean that respondents are *not* alarmed by the symbol. This is a questionable interpretation of these responses. Signs that convey a “warning,” a “danger,” or that “caution” should be used, may reasonably be expected to also cause some degree of alarm over the possible danger involved. It is not straightforward – and perhaps unreasonable – to conclude that the symbol does not cause alarm based on respondents choosing from responses 1-4.

The responses 5-7 are qualitatively different from – and incommensurate with – responses 1-4. These responses are potentially valid responses to a different question, but are inappropriate responses to the question, “What does the triangular yellow/B&W symbol mean to you?”

Response 5 refers to a respondent’s cognitive state rather than the meaning of the symbol; “it confuses me.” Another item that might have been offered along these lines might be, “it makes me less confused.” Response 6 refers to an impact on the respondent’s emotional state; “it scares me.” An additional emotional impact might be “it reassures me,” “it comforts me,” or, considering the stated goal of the question, a straightforward “it alarms me.” And Response 7 refers to an impact on the respondents’ general attention/awareness; “it gets my attention.” Another item that might have been offered might be, “it does not get my attention.”

Responses 5-7 are inappropriate responses to the question asked, ensuring that most respondents will not choose any of these items, as the question itself suggests to respondents that they should choose one of the responses 1-4.

From responses 5-7, only one of them is taken by researchers to indicate alarm. The response options are thus unbalanced in addition to being inappropriate, biasing the results toward the more numerous response types – those interpreted by the researchers as indicating an absence of alarm (Hershey et. al., 1984).

Of the seven explicit response options given to respondents, only responses 1-4 are appropriate answers to the question asked. Furthermore, three of these responses may be interpreted as indicating *some* level of alarm when viewing the symbol, and yet the researchers assert without evidence or any justification that it *indicates* an absence of alarm.

Self-Report Bias

Four sets of questions on the survey suffer from *self-report bias*. Self-report bias occurs when respondents answer questions “inaccurately” – their answers do not correspond with past, present, or future reality about them. Self-report bias reduces the survey instrument’s validity – how likely it is that the survey instrument measures what the researchers intend (for an example involving self-prediction, see Rogers and Aida, 2013).

- Respondents have limited capacity or ability for *introspection*. This means that when asked questions about their *present* selves, respondents often do not know the accurate answers — the answers that correspond with *present* reality. In questions 8 through 15 (p. 65-66) and questions 16 through 22 (p. 67) researchers ask “which sign is more helpful?” and on questions 30 through 36 they ask respondents to indicate whether “the inclusion of the specific chemical/s in the sign [...] make the sign more helpful”(p. 69-70).
- Respondents likely do not know which signs are more helpful to them *objectively* so their answers to these questions are likely inaccurate – the answers do not correspond with reality.

- The researchers should not have *asked* respondents which signs they thought more helpful but *observed which ones were more helpful*. Doing this requires an experiment/trial that assigns respondents to look at different signs and then determines which signs are more helpful by observing differences in respondents' comprehension, recall, etc. which are less abstract and more concrete measures of helpfulness.
- Respondents have limited capacity or ability at *self-prediction*. This means that when asked questions about their *future* selves, respondents often do not know the accurate answers—answers that correspond with *future* reality.
- In questions 30 through 36 researchers ask respondents to indicate whether “the inclusion of the specific chemical/s in the sign [...] help me better able to make an informed choice, make me want to seek more information” and in questions 39 through 45 ask “If you wanted additional information, how likely are you to visit the website listed below?” (p. 71). Respondents likely do not know – *they cannot predict* – whether a particular sign will – *in the future* – help them better able to make an informed choice or make them want to seek more information or visit a particular website.
- Again, the researchers should not have *asked* respondents which signs would affect their future behavior but rather *observed* which signs actually did affect behavior. Doing so requires an experiment/trial that assigns respondents to look at different signs and then determines which signs induce them to make informed choices, seek more information, or visit a particular website.

Respondent Fatigue

The study's survey instrument *fatigues* respondents not just because of its *length* but also because of the *repetitiveness* of the questions (Bradley and Daly, 1994).

Questions 8 through 15 are *exactly* the same but with different pairs of signs presented; this is true of the four additional question batteries formed by questions 16 through 22, questions 22 through 29, questions 30 through 36, and questions 39 through 45.

This means that starting with question 8, except for questions 37 and 38, respondents answer seven questions in a row with exactly the same wording within each battery, and they complete five batteries in a row like this on the survey. This *repetition* – and lack of novel stimuli – is likely to cause a decline in respondents' cognitive engagement as they progress through the survey. This decline in engagement changes how respondent will answer the researchers' questions – a form of “repetition” priming. This respondent fatigue reduces the survey instrument's validity – how likely it is that the survey instrument measures what the researchers intend.

Data Analysis

The statistical analysis of the survey results appears to have been conducted according to general practice and there do not appear to be any significant concerns in this area. The problem, however, is that the analysis necessary relies on compromised data. Statistical techniques cannot overcome the deficiencies inherent to the collected data.

A Better Methodology for Determining Label Effectiveness

The main question at issue in this report is whether or not new labeling standards are more effective than the status quo at informing citizens of potential environmental hazards. The study, however as designed and executed, is not able to answer this research question.

The study's research question is a question of behavioral responses and not a question of mere opinion. Research in many fields has shown that this distinction is crucial — that people's *actual* behavior in the real world versus what survey respondents *tell* researchers they will do in the real world -- diverge substantially.

Only randomized-controlled clinical trials can determine how an intervention, in this case signage and corresponding language, affects behavioral responses (Druckman et. al., 2011). The researchers should not have *asked* respondents which signs they thought more helpful but *observed* which ones were more helpful in fact. Determining the effectiveness of any warning label requires a randomized-controlled trial that assigns respondents to look at different labels and then measures which signs are more helpful by observing differences in respondents' comprehension, recall, etc. – less abstract and more concrete measures of helpfulness – across these different groups of respondents.

As noted above, there is a well-recognized divergence between stated intentions or evaluations and real-world behavior or impacts on behavior. In short, most individuals are not good at introspection or self-prediction. This is particularly true with respect to socially desirable behavior, such as voting, seeking out information about certain topics, or attending to public health warnings. Individuals typically overestimate or simply misreport how likely they are to take socially desirable actions.

In the case of new warning labels, we cannot determine their effectiveness at informing the public, avoiding undue alarm, or reducing overall risk without a randomized-controlled trial testing the impact of various labels on these outcomes.

The goals of Proposition 65 are numerous. And there are numerous research designs using randomized-controlled trials – clinical trials – that would more accurately illuminate key facts about how the current and proposed labels impact the public's behavior.

“Lab” and “Field” Trials

Social scientists often refer to “lab” and “field” experiments that test the impact of some stimulus or intervention on behavior. In a “lab” test, researchers expose subjects to a stimulus under conditions controlled by the researchers, in a controlled location – like an academic building – or in a controlled online environment – like an online survey – where they have greater control over the consistent delivery of the stimulus to subjects.

The impact of Proposition 65 labels and warnings can be tested in both the “lab” and the “field” – and each approach has advantages and disadvantages.

By way of example, research subjects in a “lab” experiment might take an online survey presented to them as product marketing research. During the survey, the individuals are randomly assigned to be presented with either old labels or new labels on the products they are evaluating (the experiment can be designed to tease out the impacts of various features of the labeling as well, such as the proposed icon, chemical listings, etc.). This random assignment makes the research a randomized-controlled trial – a clinical trial, the gold standard in scientific research.

The survey would task respondents with rating various features of consumer products, express their likelihood of buying each, etc., and give them the ability to actually “spend” money given to them within the survey to buy a product. The survey would also ask respondents to rate their level of anxiety and other emotional or cognitive states – to identify any impacts on respondents in these areas – and would ask factual questions about information that the labels are meant to convey.

Since the survey randomly assigns respondents to view different types of labeling, we know that any differences in the observed outcomes – and their answers – derive from the different labels and nothing else. This is the only reliable way to ascertain the real impact of proposed changes to labeling.

Following an exploration of the impacts in a “lab” setting, one might pursue further research in the “field.”

A survey such as the one conducted for the study examined here, even if designed and executed properly, cannot adjudicate the relative effectiveness of new versus old warnings. Only randomized-controlled clinical trials can determine how an intervention affects behavioral responses (Druckman et. al., 2011).

REFERENCES

Mark Bradley and Andrew Daly. (1994). Use of the logit scaling approach to test for rank-order and fatigue effects in state preference data. *Transportation*, 21:167-184.

David P. Crowne and Douglas Marlowe. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, 24: 349-354.

James N. Druckman, Donald P. Green, James H. Kuklinski, and Arthur Lupia, eds. *Cambridge Handbook of Experimental Political Science*. Cambridge University Press, 2011.

Hershey H. Friedman, Yonah Wilamowsky, and Linda W. Friedman. (1981). A comparison of balanced and unbalanced rating scales. *The Mid-Atlantic Journal of Business*, 19: 1-7.

William Kruskal and Frederick Mosteller. (1980). Representative sampling, IV: the history of the concept in statistics, 1895 - 1939. *International Statistical Review*, 48: 169–195.

Henry A. Landsberger. *Hawthorne Revisited*. Cornell University, 1958.

Elton Mayo. *Hawthorne and the Western Electric Company: The Social Problems of an Industrial Civilisation*. Routledge, 1949.

David E. Meyer and Robert.W. Schvaneveldt. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, 90: 227–23.

H.M. Parsons. (1974). What happened at Hawthorne?: New evidence suggests the Hawthorne effect resulted from operant reinforcement contingencies. *Science*, 183: 922–932.

Rogers, Todd and Masahiko Aida. (2013). Voter self-prediction hardly predicts who will vote, and is (misleadingly) biased. *American Politics Research*, 42: 508-528.

D. Steele-Johnson, Russell S. Beauregard, Paul B. Hoover, and Aaron M. Schmidt. (2000). Goal orientation and task demand effects on motivation, affect, and performance. *The Journal of Applied Psychology*, 85: 724–738.

ABOUT THE RESEARCHERS

Adam B. Schaeffer

Adam Schaeffer is Chief Science Officer at and founder of Evolving Strategies. Adam has spent the last ten years running sophisticated randomized-controlled trials in the field and in the “lab” to maximize the impact of communications.

In just the last two years alone, he led the design, execution and analysis of 22 randomized-controlled trials, testing 77 different messages across 7 states and nationally in races for governor, U.S. House, and U.S. Senate in mid-term, off-year, and special elections. The result — persuasive message impacts predicted for more than 30 million voters and tens of thousands of votes won for his clients.

Adam received his PhD from the University of Virginia in political psychology and behavior. His dissertation assessed how different combinations of school choice policies and messages can expand and mobilize elite and mass support. He received his MA in Social Science from the University of Chicago, where his thesis integrated aspects of evolutionary theory and psychology with political theory and strategy.

Adam considers himself akin to a research biologist who happens to have the great privilege of studying the behavior of the most complex and fascinating animal on the planet; *Homo sapiens*.

Alexander J. Oliver

Alexander J. Oliver is Chief Data Scientist at Evolving Strategies, a clinical data science firm that specializes in predicting and modifying human political behavior using experiments, big data analytics, and machine learning.

He's also co-founder of the firm's corporate practice, ES Partners. His recent work with the firm has been featured in *The New Yorker*, *The Washington Post*, and *The Hill*, among other media. During the 2014 midterm election season, he was on a team that spearheaded the firm's design, execution, and analysis of one of the largest field experiments in the history of political campaigning, involving the delivery of over one million messages. He primarily codes in R and Python.

He recently co-authored the definitive review article on the politics of disaster relief for the Emerging Trends project, about which best-selling authors and scientists Daniel J. Levitin and Steven Pinker have said there is "no better source of insider information on the new ideas in store for us" and have called an "indispensable reference work for the 21st century," respectively. He's taught upper-level undergraduate and graduate courses at Boston University and Brandeis University using his own state-of-the-art syllabi on campaign strategy, voter behavior, public opinion, legislative behavior, and the American Congress.

He earned an MA in Economics from Tufts University, where he was a Henken Family Scholar. He also earned a BA in Economics and Mathematics from Merrimack College, where he graduated as the top-ranked student by grade point average. He will receive his PhD from Boston University in quantitative methods and public opinion in 2016.