Analysis Plan
Project Name: Testing Variations of Maternal Immunization Messages
Project Code: 1736
Date Finalized: 11/6/2017

This document serves as a basis for distinguishing between planned (confirmatory) analysis and any unplanned (exploratory) analysis that might be conducted on project data. This is crucial to ensuring that results of statistical tests will be properly interpreted and reported. In order that the Analysis Plan fulfill this purpose, it is essential that it be finalized and date-stamped before we begin looking at the data — ideally, before we take possession of the data. Once this plan is finalized, a date is entered above, and the document is posted publicly on our team website.

Analysis Plan

Key tests

Primary Dependent Variable (DV):
The data source for the primary outcomes is ad click-through rates as provided by the marketing vendor. We will receive totals only, not raw data (which the social media platform does not release).

We will use two measures for the primary DV, based on data provided by the marketing vendor:

1) **Total number of clicks divided by total number of impressions, by condition.** Impressions are the number of times an ad has been viewed, counting multiple views per person.
2) **Total number of clicks divided by total reach, by condition.** Reach is the number of people that have been served an ad.

Secondary DVs:
The data source for the secondary outcomes is from web analytics from the marketing vendor (via the Agency collaborator) and from a web analytics page (set up by OES). These outcomes will be calculated as a rate:

- Unique pageviews (per user)
- Time on page (per user)
- Bounce rate (the number of people who navigate away from the page, after viewing only that page) (per user)
- Total events (per user)
- Unique events (per user)
- Unique event: add to calendar (from call-to-action widget) (per user)
- Total social actions (per user)
- Unique social actions (per user)

For each primary DV, the following analyses will be performed:

The proportion for each of the primary outcome variables in Condition 1 (baby-only) will be compared to the corresponding proportion in Condition 4 (mom-and-baby only). This is of the most interest to the Agency collaborator. We will use the following formula:

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

The proportions for each of the primary outcome variables will also be compared across the two core elements, images and texts. For the comparison of images, we will pool Conditions 1+2 and compare to Conditions 3+4. For the comparison of test, we will pool Conditions 1+3 and compare to Conditions 2+4. We will use the same formula as above in the comparisons across the two core elements.

Additional hypotheses will be tested as exploratory hypotheses to inform future testing (and will not be corrected). For the exploratory analysis, we will perform additional pairwise comparisons between each cell: 1 = 2; 1 = 3; 2 = 3; 2 = 4; 3 = 4.

Anticipated limitations:
Using the social media platform as the delivery mechanism reduces our ability to see process variables such as zip codes of clickers, detailed timing of ad delivery and other interesting variables, as the social media platform does not release this information. Therefore, the use of covariates and moderators is severely limited throughout the study.

We will be able to observe some covariates for the secondary outcomes (the ones captured through the web analytics); however, in the absence of information on how many adverts of each condition were delivered to the groups of people identified through this data, any analysis of this data will need to be interpreted as correlational.

It is likely that we will receive only count data for all outcomes, and no individual-level data. Therefore, all data analysis will be conducted simply by constructing test statistics for binary outcomes (unique opened email, unique clicked on email, initiated enrollment, concluded enrollment).

Specifically, we are considering including “timing of ad” (week of the season) as a covariate in the primary analysis, to account for seasonal variation or variation in how intensely the ad was displayed during any given week. However, we can only include this if we are able to get reliable indicators from the social media platform about how many ads were served up in each condition per week.

We are also considering including population size (within zip code) as a covariate. However, the data we get back from the marketing vendor is very unlikely to include zip code breakdowns of the clickers. We could try to identify the zip codes of incoming clicks through web analytics, but we will not know the denominator, i.e. how many times people in each zip code saw the advert or how many eligible individuals there were per zip code. Therefore, we can't reasonably include this control for the primary analysis. For the web analytics-based analysis, we might be able to include a control for zip, as a control for "how many people live in your zip code", and not “how many people viewed the ad campaign” itself.

Actual analysis in December 2017:
We did not receive data on clicks per zip code from the social media platform and marketing vendor, and so we were unable to control for population size in the primary analysis.

The number of web page views was not high enough to perform meaningful analysis on sub-groups of viewers, for the secondary analysis.