B. STATISTICAL METHODS

1. Universe and Respondent Selection

The research will be conducted using non-probability Internet panels maintained by Ipsos which have been designed to be representative of the adult population. Additional information about the universe and respondent selection is provided in the attached supporting document, "Research on Consumer Tipping Behavior: Response to OMB Information Request." This document further elaborates on sample balancing, quality assurance processes, estimation procedures, outreach/advertising/recruitment methods, and existing research on panel comparison to benchmarks.

2. Procedures for Collecting Information

Potential respondents will be sent a survey invitation inviting them to join the survey effort, at which point they will be asked for relevant demographic information prior to accessing the survey.

3. Methods to Maximize Response

We are utilizing an established online panel for survey administration. Survey administration will include an invitation email and up to one reminder email (as needed) in an effort to maximize response rate.

4. Testing of Procedures

Prior to determining the use of the online panel for the full-year survey fielding, FMG conducted a one-month pilot study to arbitrate between two pilot samples. This pilot study was conducted according to OMB guidelines for deciding between two possible samples. The pilot study compared the bias in the estimated mean tipping rates derived from responses taken from the non-probability online panel and a probability-based push-to-web panel. The full report detailing the pilot study and outcomes, "Comparison of Estimates of Tipping Behavior Produced Using Probability and Non-Probability Samples: Methodology and Results," is attached as a supporting document; the key findings are discussed in this section.

The pilot data analysis featured two tests of the relative bias in the two estimates. The first test, termed the "Differences in Samples" test, assumed that the probability sample is no more biased than the non-probability sample. Consequently, any difference in reported average tip rates between the two samples was interpreted as indicating bias in the non-probability sample. The results of this test found no statistically significant differences between the mean tipping rates derived from the two samples.

The second, "Differences in Differences" test did not make an assumption that the probability-derived estimate was not more biased than the non-probability estimate of the mean tipping rate. Rather, this test utilized information about tipping transactions from point of sale data

(POS) as an objective arbiter between the probability and non-probability samples. Specifically, the test examined whether the absolute mean difference between respondent-reported tip rates and the mean tip rates of the respondent's region of residence differed between the non-probability and probability samples. This test found no evidence that the non-probability estimate systematically differed from the POS estimate more than the probability estimate.

As with any study, this analysis was subject to some limitations. Specifically, the pilot study compared tip rates between the two samples for only a subset of tipped transactions of interest (Full Service Restaurants), so it is unclear to what extent the findings generalize to non-restaurant transactions. In addition, neither the probability-based estimate nor the Point of Sale data represent a gold standard even for restaurant tips. The Point of Sale data was taken for a non-randomized sample of establishments, focused solely on credit-based transactions, and thus might not be comparable to the sample of transactions taken from the survey, which included all forms of payment, including cash. Finally, non-response in both panels means final estimates from either sample will come with a level of uncertainty concerning the degree to which they represent the "true" mean tipping rate. However, the pilot study does not provide systematic support for the superior validity of either the probability or non-probability sample.

Although the results of neither test clearly supported one sample being more biased than the other, the results supported the use of the non-probability sample. Specifically, given considerations of the cost of obtaining a sample of sufficient size to produce estimates not just for full-service restaurants, but for other, more infrequent tipping industries as well as the robust lack of evidence for a difference in the bias in the estimates of the mean tipping rate, the non-probability sample was deemed preferable.

The primary criterion for determining the minimum target sample size for the full yearlong survey was the ability to produce valid estimates for the national mean tip rate for each industry with a margin of error not exceeding 2 percentage points. Other criteria, such as the precision for analyses of seasonal effects or geographic differences among more frequently tipped industries, are of secondary importance and were not under consideration when determining minimum sample sizes.

In order to meet the desired level of precision, it was determined necessary to have 1200 tipping occurrences per industry over the course of a year, or to average 100 tipping occurrences per month for each industry. The table below shows the estimated number of completed surveys needed to produce a national mean tip rate with the desired level of precision for each industry (shown in final column). These estimates were determined using incidence rates of voluntarily tipped occasions obtained during the pilot study from the Ipsos, non-probability sample. These incidence rates are higher than indicated in other sources, and thus will result in data for more tipping incidences for a given sample. However, given that the pilot study was undertaken for one summer month, the incidence rate may not be representative of what would be obtained from a yearlong fielding period. Consequently, to be conservative, the incidence rates in the tables should be interpreted as an upper bound, particularly for transaction types such as "Hotel/motel" and "Casino" which are likely to display substantial seasonal variation. As shown, the industry with the lowest incidence is "Moving or household maintenance services" and the number of completed survey responses necessary to produce a mean tip rate for that industry with a margin of error of 2 percentage points or less is 57,143. This has determined our target sample size for the full yearlong survey, which is 60,000 completed responses.

Estimated Annual Tipped Occurrence – Ipsos Pilot Study Data (N=7,050)

	Occasions per year	Likelihood per day	Required sample for 1,200
Restaurant or other prepared food/drink service	120.5	**0.330**	**3,636**
Hotel/motel	10.6	**0.029**	**41,379**
Personal grooming, beauty, or massage services	35.0	**0.096**	**12,500**
Moving or household maintenance services	7.7	**0.021**	**57,143**
Casino	12.0	**0.033**	**36,364**
Taxi, limousine, rideshare, or shuttle service	13.1	**0.036**	**33,333**

To mitigate potential bias in the mean tipping rate, a post-stratification procedure will be applied to the full fielding data such that the weighted sample of respondents will match the adult U.S. population with respect to gender, age, education, race/ethnicity, income, the fraction of the respondent's county which is foreign born, the rural/urban status of the respondent's county, and census division. To account for any effects the date of the transaction might have on both tips and response rates, and thus on estimate bias, the post-stratification procedure may also give larger weight to respondents whose period of recall takes place during months and/or days of the week which are underrepresented in the survey. This weighted sample of respondents will then be weighted by the respondent's expected number of tipped transactions such that the mean tipped rate is representative of all tipped transactions.

Prior to the pilot study, FMG conducted a usability study with 35 adults prior to the launch of the pilot study. Testing occurred in November and December of 2014. These participants tested the survey language (to ensure survey respondents understand the industry/service) as well as tipping (monetary/in-kind) attribute language and can accurately recall their tipping activity. The survey findings indicated that some minimal wording changes were required. The IRS Office of Research and FMG did not find that any significant changes were necessary based on the findings of the pilot study. Such potential changes that were considered included an increase in the kind or amount of information sought; an increase in coverage; an increase in the timing or frequency of reporting; a change in the sample design or collection method; or a change in the purpose for which the information is collected or required to be maintained.

FMG reviewed the findings from the pilot study and determined there was no compelling evidence to institute a change in the period of recall time. In addition to the pilot survey findings, the usability testing included a review of possible differences in recall time used for the survey and led to the conclusion that the recall period should be 1 day for all transactions. One of FMG's findings were that respondents appeared to more heavily lean on the use of estimation heuristics as the recall period was lengthened from 1 to 3 to 5 days (e.g., "It wouldn't really be that difficult for me to recall [longer period of time] since I usually tip about 15%"). More information on the findings from FMG's cognitive and usability testing on this project can be found in the full report highlighting findings and edits made to the survey instrument.[1]

[1] *IRS Tipping Report on Cognitive and Usability Testing*, January 2015. Internal report prepared for the Internal Revenue Service

A recall period longer than 1 day could also lead to other forms of recall errors. One such type of error that has been found in research concerning major events or large consumer purchases is telescoping, which occurs when the respondent fails to accurately remember when an irregular event occurs or even “remembers” events that never occurred. This could be very problematic during analysis as it would be impossible to determine the rate at which these services are occurring and on what days. Increasing the number of observations of certain, less frequent services would be ideal, but not at the expense of significantly reduced data quality per observation that could result from these recall errors or reliance on estimation heuristics.

Lengthening the recall period beyond 1 day would also increase respondent burden and could lead to reporting bias. Respondents would have to carefully read instructions for each series of questions to determine when different recall lengths are being asked, and then would have to remember specifically which day the expenditure occurred in order for it to be useful for analysis. Such a significant change in survey instructions would require additional testing to ensure comprehension. Confusion could also result from determining which was the most recent expenditure, and could allow for reporting biases concerning which expenditure they choose to report. This would be particularly problematic for service categories where multiple expenditures could occur at the one establishment (like hotels, beauty salons, or casinos). In such scenarios, respondents might choose to report a service that they can recall easily rather than reporting the one that they paid for most recently.

The survey will be administered electronically; however there are no cookies involved. Survey participants will be provided a link/web address via a secure website. Transmission to/from the secure website for the survey will be encrypted.

Survey respondents will be selected from the subcontractor’s panel members. Participants will be provided a link/web address to a secure website with their unique survey URL that corresponds to their survey questions. The subcontractor hosting the panel and survey will maintain a secure survey control system that will document the correspondence and track the status of all sample members by giving each sample member a unique sample ID. The sample ID is used in place of name, address, or other personally identifiable information.

5. Contacts for Statistical Aspects and Data Collection

For questions regarding the study or questionnaire design or statistical methodology, contact:
Brian K. Griepentrog, Ph.D.
Director of Research Studies
Fors Marsh Group LLC
bg@forsmarshgroup.com

by Fors Marsh Group under contract TIRNO-13-Z-00021-0002.