Description of document: Internal Revenue Service (IRS) records relating to contract Research on Consumer Tipping Behavior, 2014-2018

Requested date: 28-November-2018

Release date: 10-April-2019

Posted date: 17-June-2019

Source of document: IRS FOIA Request
HQ FOIA
Central Processing Unit
Stop 211
PO Box 621506
Atlanta, GA 30362-3006
Fax: 877-807-9215

DEPARTMENT OF THE TREASURY
INTERNAL REVENUE SERVICE
WASHINGTON, DC 20224

PRIVACY, GOVERNMENTAL
LIAISON AND DISCLOSURE

April 10, 2019

This is our final response to your Freedom of Information Act (FOIA) request dated November 28, 2018 that we received on December 11, 2018.

You asked for a copy of the Statement of Work, the interim reports/presentations, the final report/presentation, and successful proposal for contract GS00F348CA awarded to FORS Marsh Group Ltd. to study tipping behaviors. Of the 241 pages located in response to your request, I am enclosing 220 pages. I am withholding 21 pages in full for the following reason:

I am withholding the successful proposal in full under FOIA exemption (b)(3). This exemption requires us to withhold information that is specifically exempted from disclosure by another law. The law supporting this exemption is Title 41 United States Code section 4702.

I am providing your documents on the enclosed encrypted CD. The password to open the CD is being sent separately. This constitutes a partial denial of your request. There is no fee for processing your request.

You may contact me, the FOIA Public Liaison, Laura A. McIntyre, to discuss your request at: IRS Disclosure Office 10, 1999 Broadway, MS 7000, DEN, Denver, CO 80202, (303) 603-4734.

The FOIA Public Liaison responds to FOIA and Privacy Act requests for copies of documents maintained by the IRS. There is no provision in either Act to resolve tax, collection, or processing issues and our staff is not trained to answer questions regarding those issues. If you need assistance with tax related issues you may call the IRS toll free number at 1-800-829-1040.

If you are unable to resolve your FOIA dispute through the FOIA Public Liaison, the Office of Government Information Services (OGIS), the Federal FOIA Ombudsman's Office, offers mediation services to help resolve disputes between FOIA requesters and Federal agencies. The contact information for OGIS is:

Office of Government Information Services
National Archives and Records Administration
8601 Adelphi Road--OGIS
College Park, MD 20740-6001
202-741-5770
877-684-6448
ogis@nara.gov
ogis.archives.gov

You have the right to file an administrative appeal within 90 days of the date of this letter. By filing an appeal, you preserve your rights under FOIA and give the agency a chance to review and reconsider your request and the agency's decision. I've enclosed Notice 393, *Information on an IRS Determination to Withhold Records Exempt From the Freedom of Information Act - 5 U.S.C. 552,* to explain your appeal rights.

If you have any questions please call Disclosure Tax Law Specialist John M Quigley ID # 1000247712, at 510-907-5337 or write to: Internal Revenue Service, Centralized Processing Unit – PO Box 621506, Stop 211, Atlanta, GA 30362. Please refer to case number F19345-0060.

Sincerely,

Laura A McIntyre
Disclosure Manager
Disclosure Office 10

Enclosure
  Responsive Records
  Notice 393

RESEARCH ON CONSUMER TIPPING BEHAVIOR

Task Order 4

One Year of Data Collection

1. Background

Task Order 4 begins the third phase of a multi-year project to collect data on consumer tipping in order to estimate total tips paid and to better understand tip income reporting compliance.[1] The data collected from this project will be used to estimate total tip income as well as tipping and stiffing rates by method of payment (e.g. cash, credit card, and debit card) and by industry (e.g. restaurants, hotels, casinos, taxis, barber and beauty salons).[2]

In Task Order 1, the contractor Fors Marsh Group (FMG) conducted a review of the literature on tipping behavior, identified and discussed options for key study elements (i.e. sampling source, sampling mode, study design, and data analysis), and recommended a plan of analysis. The analysis plan recommended a repeated cross-sectional design with an internet-based panel sample and survey questionnaire. In Task Order 2, FMG designed, tested and refined the survey questionnaire. For Task Order 3, FMG is conducting a one-month pilot of the web-based questionnaire developed in Task Order 2 to determine which of two sampling strategies – probability or non-probability – offers the best combination of quality and cost.

In Task Order 4, FMG will conduct one year of survey data collection of consumer tipping behavior. The data collection will follow the recommended analysis plan developed in Task Order 1, using the survey questionnaire developed in Task Order 2 and the sampling strategy determined by Task Order 3. In addition to overseeing and monitoring the data collection, performing data validation, and providing the survey data to IRS, FMG will also prepare a technical report describing the survey methodology and implementation. Optionally, FMG will produce sample weights for the survey data, conduct statistical analysis to generate aggregate measures of tipping behavior, and evaluate the possibility of merging the one-month pilot data from Task Order 3 with the full year survey data.

2. Objectives

The main objective of this project is for IRS to acquire data that can be used to produce several estimates related to tipping behavior, such as total tip income, tipping rates and stiffing rates, all by industry, major method of payment (e.g. credit card, debit card, and cash), and geographic region, where possible. These data will ultimately be used to produce aggregate estimates of underreported tip income in order to improve IRS estimates of the tax gap and to improve IRS's determination of unreported tip income during individual taxpayer audits. This project may optionally include some or all of the analysis to produce these desired estimates.

[1] Blanket Purchase Agreement (BPA) contract number TIRNO-13-Z-00021 awarded to Fors Marsh Group (FMG) on September 30, 2013.

[2] "Stiffing" occurs when a tip is expected but not paid.

3. Tasks

This project consists of both core and optional tasks.

3.1 Core Tasks

1. Administer the survey questionnaire during calendar year 2016. Under Task Order 4, FMG will submit a clearance package to OMB and obtain OMB clearance to carry out one year of data collection beginning in January 2016. The survey questionnaire has already been developed under Task Order 2. Based on the results of analysis conducted under Task Order 3, and in consultation with IRS Office of Research, FMG will determine the appropriate method of sample selection (probability vs. non-probability) and the target sample size.
2. Compile survey data into a single ready-to-use data file. FMG will receive raw survey data from the vendor, compile it as necessary and perform data validation and quality control.
3. Prepare a technical report describing the survey. FMG will prepare a report that includes a description of the sample design, methodology and implementation of the survey. The report should contain standard information to ensure that analysts can work intelligently with the survey data.

3.2 Optional Tasks

1. Produce sample weights for the survey data. FMG will generate sample weights so that aggregate measures and population-level parameters can be estimated from the survey data. A written description of the sample weighting process will also be provided.
2. Perform statistical analyses to generate summary measures of tipping behavior. Using the survey data and sample weights, FMG will perform analysis to generate national estimates of total tip expenditures, as well as subtotals of tip expenditures by industry/occupation, method of payment, and geographic region, depending on the level of detail permitted by the survey data. FMG will also produce estimates of stiffing rates and tipping percentages by industry/occupation, method of payment, geographic region, and other factors regarding the nature of the tipping occasion (e.g., type of restaurant or size of establishment) to the extent permitted by the survey data. The methodology and outcomes of the analysis will be presented in a written report.
3. Evaluate the possibility of merging the pilot data with the full survey. Depending on the number of respondents and number of tipping occasions that are captured by the full year survey, it may be beneficial to enhance the survey data with data collected during the pilot fielded in July 2015 under Task Order 3. An optional task would be to evaluate the merits of such an approach and outline methodology for doing so.

The contractor shall provide monthly status reports to the Contract Officer's Representative (COR) and IRS' Technical Lead on the progress of their work. The reporting will be in the form of an e-mail with follow up telephone communication, as needed. The contractor will be available to the COR and Technical Lead during normal business hours.

4. Deliverables

This project consists of both core and optional deliverables.

4.1 Core deliverables

1. Weekly reports during the one-year survey period with counts of total contacts with survey respondents, number of completed surveys, and number of completed surveys with a tipping occasion.
2. A flat file copy of the survey responses collected during the one-year survey period including supporting information (e.g., data dictionary, record layout). Data files will be in an industry standard format acceptable to IRS (e.g. comma or tab-separated values or SAS format).
3. A technical report describing the survey methodology and implementation.

4.2 Optional deliverables

1. Sample weights created by FMG included in a flat file and described in a written report.
2. A written report containing estimates of tipping behavior and methodology for those estimates.
3. A report describing the technical considerations and merits of merging pilot data with survey data and a recommended methodology for doing so (if appropriate).

5. Data Safeguards

No IRS data will be used in this study and no Personally Identifiable Information will be collected for the survey respondents.

6. Timeframes

IRS expects that the core aspects of this effort will require approximately 1.5 years from the date of contract award to completion. Optional tasks and deliverables may take several months longer and timeframes may depend on which of the optional tasks are undertaken. The IRS proposes the following set of milestones and deadlines:

Milestone	***Description***	***Due Date***
1.0	Kickoff Meeting	5 days from contract award
1.1	Submit OMB package for Field Survey	15 days from contract award
1.2	Begin collecting data	January 1, 2016
1.3	Complete collecting data	December 31, 2016
1.4	Deliver draft technical report and core data set (optionally including sample weights and supporting documentation)	March 15, 2017
1.5	Deliver final technical report and finalized core data set	April 30, 2017

	(optionally including sample weights and supporting documentation)	
1.6	(Optional) Deliver draft report containing estimates of tipping behavior	July 15, 2017
1.7	(Optional) Deliver final report containing estimates of tipping behavior	August 15, 2017
1.8	(Optional) Deliver draft evaluation of merging pilot and full survey data	Contingent on other optional tasks
1.9	(Optional) Deliver final evaluation of merging pilot and full survey data	Contingent on other optional tasks

Alternative timeframes will be permitted upon IRS approval.

7. Ownership of Research Findings

The IRS will own the results of this study including the raw survey data, any supplemental data such as sample weights or other written information or documentation generated by the contractor. The IRS and the Office of Research understand the importance of publishing tax administration research in professional publications and supports and encourages such activity. However, IRS review and written approval must be obtained before presenting or publishing any study based on the results of this research.

8. References

Estimating Consumer Tipping Behavior: Review and Recommendations. Final report prepared by Fors Marsh Group, LLC under IRS contract TIRNO-13-Z-00021.0001. February 2014.

IRS Tipping Report on Cognitive and Usability Testing. Final report prepared by Fors Marsh Group, LLC under IRS contract TIRNO-13-Z-00021.0002, January 2015.

Comparison of Estimates of Tipping Behavior Produced using Probability and Non-Probability Samples. Report prepared by Fors Marsh Group, LLC under IRS contract TIRNO-13-Z-00021.0002, January 2015.

FORS|MARSH
GROUP

Estimating Consumer Tipping Behavior: Review and Recommendations

Prepared for: Internal Revenue Service

Prepared by: Fors Marsh Group, LLC

FINAL

Version 2.0

February 2014

Table of Contents

Introduction

This report is intended to provide guidance to the IRS as it attempts to develop estimates of tipping and stiffing rates, tipping income, and ultimately, the gap between actual and reported tip income at the aggregate level and by sector. This guidance is based on the results of past research on tipping behavior as well as lessons learned from authors' own work in this area. The first section of this report summarizes the results from a comprehensive annotated bibliography of academic and government literature on tipping. This bibliography, which can be found in Appendix B, includes summaries of research examining average tipping rates as well as individual and establishment characteristics associated with tipping. In anticipation of subsequent sections, the bibliography also summarizes articles that do not directly address tipping, but are relevant to the development of research designs that could be used to collect and analyze data on tipping.

The second section reviews different methods for the collection and analysis of tipping data, and their potential benefits and drawbacks. Topics addressed include sample sources, specifically samples drawn from address-, telephone-, and Internet-based samples; the mode used to collect the data from the sample, including in-person interviews, paper surveys, and Internet surveys; and the design of the survey, including long-recall cross-sectional, short-recall repeated cross-sectional, and longitudinal designs. Finally, this section describes potential methods for analyzing the data, including the use of disaggregated means as well as model-based approaches.

Finally, the third section presents recommended approaches for collecting and analyzing tipping data based on the reviews in the first two sections. This includes both immediate steps pertaining to instrument development as well as pilot testing prior to full scale implementation.

Literature Review Summary

A preliminary set of articles was identified using a bibliography of tipping -related research compiled by Dr. Michael Lynn. [1] Additional articles were identified through backward and forward citation searches starting from the articles identified in the Lynn bibliography. Google Scholar was used to identify more recent research that cited the articles from Lynn's bibliography. Gated articles were accessed through a local University Library System. However, to mitigate the potential for selection bias, queries for articles relevant to tipping and survey methodologies were made using several search engines and archives. This set of search engines and databases included general interest academic archives and search engines such as Google Scholar, JSTOR, and the Social Science Research Network (S SRN) as well as specialized business and accounting -related archives such as Business Source Complete and ProQuest's Accounting & Tax database. Themes and keywords for this search were identified based on an initial review of articles obtained from the Lyn n bibliography and the backward and forward searches. From these articles, additional backward and forward searches were conducted to identify additional articles. From the resulting compilation of articles, authors influential to the tipping literature were identified based on total numbers of articles written/ published and/ or number of citations. These researchers were consulted in order to obtain any previously unidentified tipping -related papers/ research, whether published or unpublished.

Many articles touch on multiple topics that are relevant to determining a methodology for data collection and analysis of tip -stiffing and tip rates. Consequently, articles cannot be sorted into mutually exclusive categories based on themes. To facilitate review of evi dence from the compiled literature on specific topics, each citation includes a list of the article's themes. The reader can use his or her word processor/ PDF reader's search or find functions to quickly discover articles that address a given theme. A list of all themes with descriptive text is included in Table 1. A list of the reviewed articles is provided in Appendix A, with the associated annotations presented in Appendix B. Descriptions of search engines, search terms, and related themes derived from t he search are provided in Tables 2 and 3 in Appendix C.

[1] http://tippingresearch.com/uploads/Tip_Bibliography.pdf

Table 1. Themes

Theme	Description
METHODOLOGY	Methodology used by article along with relevant benefits and drawbacks.
NATIONAL AVERAGE TIPPING RATES	Article's findings, if any, with respect to U.S.-wide stiffing/ tipping rates.
INDUSTRY/ SERVICE	Article's findings, if any, with respect to differences in average stiffing/ tipping rates across industries/ establishment types.
CASH VERSUS CREDIT	Article's findings, if any, with respect to differences in stiffing/ tipping rates between establishments/ customers who accept/ use cash versus credit.
SERVICE CHARGE	Article's findings, if any, with respect to differences in stiffing/ tipping rates between establishments that do or do not include automatic tip/ service charge.
BILL SIZE	Article's findings, if any, with respect to differences in stiffing/ tipping rates between establishments/ customers based on bill size.
GEOGRAPHY	Article's findings, if any, with respect to differences in average stiffing/ tipping rates across geographic regions and jurisdictions.
INCOME	Article's findings, if any, with respect to differences in average stiffing/ tipping rates of customers with different levels of income.
EDUCATION	Article's findings, if any, with respect to differences in average stiffing/ tipping rates of customers with different levels of educational attainment.
AGE	Article's findings, if any, with respect to differences in average stiffing/ tipping rates between customers based on AGE.
GENDER	Article's findings, if any, with respect to differences in average stiffing/ tipping rates between men and women.
RACE/ ETHNICITY	Article's findings, if any, with respect to differences in average stiffing/ tipping rates of customers with racial/ ethnic characteristics.
TIPPING KNOWLEDGE	Article's findings, if any, with respect to customers' understanding of tipping norms (i.e., percent of bill).

Methodology: With respect to the methodologies related to collecting data on consumer/ producer expenditure and reporting, the current literature covers many of the trade-offs between maximizing data quality, making causal inferences, and ensuring that the sample and their recorded behavior is representative of the population of interest. Panel-based survey designs, such as the original NPD Group diary panel (McCrohan & Pearl, 1991; Pearl & McCrohan, 1984) and the Bureau of Labor Statistics' Consumer Expenditure Survey, can potentially allow analysts to make inferences about the effects of interventions on individual behavior because of the ability to control for individual-level factors that do not vary over time (Parker, Souleles, & Carroll, 2012). However, panel-based survey designs can also potentially increase respondent burden, leading to increased attrition and selection bias. In addition, consumer diary panels may induce changes in respondent spending behavior, leading to less valid predictions for individuals outside the sample (Crossley & Winter, 2012). A similar trade-off comes with experiments, whether in labs (Alm & Jacobsen, 2007) or fields (List, 2011)[2], which allow for controlled environments, and thus the estimation of treatment effects at the expense of external validity. Nonpanel surveys, while providing limited ability to make causal inferences about the effects of different interventions on expenditure and reporting, can potentially produce more representative samples because of a relatively lower burden being placed on respondents and consequently higher response rates. However, long recall periods may lead to lower quality of responses because of the inability of respondents to accurately recall the timing of spending occasions (Crossley & Winter, 2012).

With respect to the effects of survey modes and instruments, web-based surveys can lead to more accurate responses than paper-based or in-person/telephone surveys because of the ability of respondents to more easily skip past irrelevant questions, and, in the case of in-person/telephone surveys, the increased time respondents have to look up information necessary to accurately answer questions. In addition, self-administered surveys, which are now primarily web-based, may be more accurate than in-person or phone interviews because of the anonymity that self-administered web-based surveys afford (Crossley & Winter, 2012). However, a sample of individuals with web access may not be perfectly representative of the population of interest because of an individual's probability of web access being related to individual characteristics as well as his or her geography.

Industry/Service: The majority of tipping research has focused on the restaurant industry, but a few studies have focused on other industries where tipping is prevalent. For instance, previous studies have investigated tipping rates for luggage handlers, taxi drivers, bartenders, parking attendants, hotel bellmen, and barber/hair stylists. Koku (2005) concluded that there is a difference between tipping rates in the restaurant industry and outside of it by interviewing customers of each sector. Similarly, Paul and Gardyn (2001) identified higher tip percentages for restaurant servers than for barbers, taxi drivers, food delivery workers, hotel bellmen, and several other professions. However, more research is necessary to provide a more direct comparison between customers' tipping behavior in restaurants and other service industries.

A relatively significant amount of research has been conducted to investigate alcohol's effect on tipping rates. Most of that research demonstrated that customers who consume alcohol provide higher tip percentages than those who do not. Even after controlling for the relationship between bill size and tip percentage and a host of other variables, Lynn (1988) identified a significant effect of alcohol consumption on tip percentage. Similarly, Bodvarsson and Gibson (1997) found that tip rates varied across establishments, but establishments that were licensed to serve alcohol received higher tips.

Cash Versus Credit: Several studies have investigated the difference in tipping rates between restaurant patrons who pay their bill with a credit card and those who pay with cash. Although some articles failed to find a significant difference between payment methods, the majority of research on this subject seems to indicate that customers who paid with credit cards tipped at higher rates than those who paid with cash.

Bill Size: Some research has focused on what is known as the magnitude effect of tipping. The magnitude effect refers to the tendency for customers to leave bigger proportional tips for smaller

[2] List, J. A. (2011). Why economists should conduct field experiments and 14 tips for pulling one off. *The Journal of Economic Perspectives*, *25*(3), 3-15.

bills compared with bigger bills. Chapman and Winquist (1998) concluded that customers provide higher tips on smaller checks, and that the tip percentage decreases as the total bill increases by demonstrating the effect in restaurants and barbershops/hair salons.

Regional Differences: Geographic difference such as between urban and metropolitan areas and between different census regions or divisions of the country, have been investigated previously in studies that contained sufficient sample sizes. Typically, customers in urban areas have been shown to tip at higher percentages than ones in rural areas, and there are some studies showing significant differences in tipping behavior and knowledge between customers from the Northeast region of the country and those in other parts of the country. For instance, McCrohan and Pearl (1983) found that Northeast customers tipped at higher rates than those from middle parts of the country, and Lynn (2006) reported that Northeast customers had a higher knowledge of tipping norms than those in the South. However, the urban/rural difference and regional could potentially be explained by differences in racial composition as well as education and income levels. Studies have demonstrated that these variables significantly influence tipping behavior and tipping knowledge, with higher income levels and educational levels leading to higher tipping and knowledge of tipping norms.

Age and Gender: have been investigated in a number of studies on tipping behavior, and the influence these variables exert on tipping behavior has been somewhat mixed. Gender differences have been inconsistent across studies, whereas age has been linked to significantly higher tipping rates (Pearl & Vidmar, 1988) and greater knowledge of tipping norms (Lynn, 2006). However, differences in tipping behavior and tipping knowledge due to age could be confounded by other factors, such as higher income levels, differences in payment methods, and educational differences. For instance, Lynn (2006) showed in his analysis that higher knowledge of tipping norms with increasing respondent age was nonsignificant when other factors such as education, income, and metro status were included in the analysis.

Race Ethnicity: One of the most researched topics in tipping compliance is how tipping behavior differ between various racial/ethnic groups. Numerous studies have researched this topic, investigating not only actual tipping behavior, but knowledge of tipping norms as well. The findings of these studies have some robust conclusions; primarily, that Black customers tip at lower rates than White customers in the restaurant industry. Though less researched, studies investigating racial/ethnic differences in other tipping industries have reached similar conclusions: Black customers usually tip at lower rates, stiff more often, and tend to leave "flat" tipping rates at higher percentages than White customers, as noted in Lynn's 2004 study on Black-White tipping differences among service industries (though this effect differs across certain service industries).

Although a significant amount of research has investigated the differences between White and Black tipping behavior, some work has looked into differences between Asian and Hispanic customers. Studies by Lynn in 2006 and 2013 on tipping rates and knowledge of tipping norms indicated that Hispanics tip at lower rates than Whites and have lower levels of tipping knowledge. Asian customers in these studies were not shown to tip at significantly different levels or be less knowledgeable about correct tipping norms. However, these racial/ethnic groups have not been as thoroughly researched as Black and White customers.

Review of Methodologies

For the purposes of developing estimates of consumer tipping by industry, multiple approaches can be taken with respect to the method used to collect the data as well the method used to analyze the resulting data. A data collection approach is defined along several dimensions; specifically, the choice of sampling source, survey mode, and survey design. Methodologies used to analyze the data can be roughly categorized into simple, nonparametric approaches and parametric, model-based approaches.

Sampling Sources

The primary factors to consider when choosing a sampling source are the representativeness of the resulting sample with respect to tipping behavior and the costs associated with recruiting and retaining the sample. Sampling-related bias can result from an unrepresentative frame, non-response, incompletes, and, in the case of a longitudinal panel, attrition. In addition, to mitigate the potential for additional nonresponse that results from transitioning sampled respondents to the survey, survey sources are often coupled with survey modes. To minimize total survey error, biases that result from the survey mode have to be considered when choosing a sampling source. Table 2 discusses the benefits and drawbacks of different sampling sources.

RDD Sample: Random digit dialing (RDD) uses randomly generated phone numbers to select a sample for participation in a survey. RDD sampling is helpful in the sense that it may allow for coverage of the population that has unlisted numbers, but there are problems associated with cell phone users. First, in some RDD methodologies, cell phone users are not reachable, which excludes individuals without household lines from the sample, affecting the generalizability of results. Second, even if cell phone users can be reached, it may not be possible to determine the participant's location from his or her area code because many cell phone users retain their old numbers when moving to different regions of the country, forcing researchers to rely on self-reporting. Self-reporting of basic demographics has a negative influence on response bias analyses because little to nothing is known about the nonrespondents. This is compounded by the fact that RDD response rates continue to decrease with the widespread adoption of caller ID and call screening.

Address-Based Sample: An address-based sampling (ABS) source relies on home address and demographic information from the frame file, which is provided by third-party vendors and the U.S. Postal Service (USPS). This source allows for measureable sample coverage across a population and a fairly well-known probability of the sample selection. Additional household data can be purchased and appended to ABS files to assist in more targeted sampling and further response bias analyses. Although the costs for a mail paper sample are not low, cost per complete has been found to be lower than that for RDD studies[3] with certain populations. However, response rates for mail-based paper surveys, which are the most commonly used data collection mode with a mail paper sample,

---

[3] Medway, R.L., Viera, L., Turner, S.R. & Marsh, S.M. (2009). *List-assisted mail as an alternative to random digit dial in a survey of the young adult population.* Paper presented at the 64th Annual Conference of the American Association for Public Opinion Research, Hollywood, FL.

have continued to decrease over time as people increasingly to use the Internet as their medium for correspondence (Dillman et al., 2009b).

Traditional In ternet Sample: Traditional Internet samples are collected via an opt -in procedure where individuals choose to join a survey administrator's panel. This panel acts as a potential pool of respondents who are then queried to participate in individual surveys or diaries. An opt-in sample might not be representative of the population of interest with tipping or expenditures on tipping services. Unlike RDD or ABS methods, randomly sampling from an email -based Internet frame is made difficult by the lack of an al gorithm that can randomly select email addresses, due to inconsistencies in email address conventions (Dillman, 2009). Yet, similar to the RDD and ABS methods, traditional Internet samples can fit with various data collection modes. Although the online survey is the most straightforward mode for Internet samples, it should be noted that a longitudinal diary approach could also be used, such that all those potential respondents are contacted and recruited to report their tipping behavior for a predetermined amount of time (see Survey Design). Some examples of different Internet samples include:

- GfK Knowledge-Panel®: The KnowledgePanel is an Internet -based panel that uses a probability-based sampling strategy where the survey frame is derived from the USPS Delivery Sequence File. Individuals are invited to participate in the panel by mail, followed by telephone calls for those who do not respond to the initial invitation. Households are sampled without replacement, avoiding potential bias that may result from r espondents participating in the panel twice. For those individuals selected for participation without computers or an Internet connection, a netbook is provided. This process attempts to mitigate the selection bias associated with web surveys while preserv ing the benefits associated with a computer interface. The primary benefit of the Knowledge Panel relative to the opt-in panels described below is that knowing the probability of selection allows researchers to estimate error. However, these estimates will always be deficient capturing all aspects of non-response unaddressed by demographic post -stratification. Further, the procedures used to setup and maintain panel membership and participation serve as an additional component of error difficult to fully mod el and correct for.

  Blended Online Sample (Ipsos Ampario): Ipsos' blended sample approach combines the use of its Ampario online sampling method in addition to its iSAY online panel—an online panel of 800,000 members and their households. Ampario is a new, nonprobability sampling procedure that Ipsos has developed that invites respondents by invitations , banner ads, and other means on 100 to 400 websites that have partnered with Ipsos. These two methods are combined into a single sample using Ipsos' pro prietary Cortex routing system, which allocates and reallocates a sample given respondent eligibility. Simply put, when respondents are not eligible for one survey, they are immediately redirected to other surveys in progress. In traditional one -off opt-in surveys, noneligible respondents are lost sample, a considerable cost. Finally, Bayesian methodology, which requires previous information regarding the overall sample of interest in order to mix with current information for the final distribution of resul ts, is used to form final distribution. As is the case with a traditional online sample, Ipsos' blended sampling could work with several different data collection modes, but it is best served with an online -based questionnaire, which could include a cross -sectional administration or a longitudinal diary approach. However, because of the opt -in nature of the Blended Sample, it is not possible to model the probability of response, and thus account for that source of potential bias in survey estimates.

- NPD Group Online Sample: NPD Group utilizes sophisticated techniques both at the sample design stage and post -survey weighting stage to reduce bias and increase representativeness of the sample, but it is not a probability sampl ing technique. . Although there are certain demographic groups that have less representation online and are not represented in correct proportions as they would be in the U.S. Census, they are large enough that they can be sampled appropriately to represent the U .S. population.

  Recruitment of panelists is done using a wide variety of opt -in sources (email, affiliate marketing, co-registration, banners , etc.). The wide variety of sourcing ensures a large representation from various strata of the U.S. population. All sourcing is b alanced and ensures no single source provides a disproportionate percentage of recruits. A number of other steps are put in place to prevent fraudulent prospects from joining the panel. This, combined with other behavioral data collected, is used to monito r recruitment source quality and guide media planning for recruitment. NPD limits the number of surveys a panelist can start in a day, week, and month to avoid survey fatigue. Response rates are tracked at an individual panelist level —if panelists fail to participate consistently over time, NPD removes them from the active panel.

  The sample for a particular study is drawn from this panel to demographically represent the U.S. population. Sophisticated algorithms take varying response rates by demographic groups into account to provide stratified quota for each of the targeted cells. Once the sample is collected , the cells that fall short in demographic representation during sampling are weighted during processing the data. Again, because of the opt -in nature, it is not possible to model the probability of response, and thus account for that source of potential bias in survey estimates.

Table 2. Summary of Trade-Offs in Sampling Sources

			INTERNET SAMPLES			
Qualities of Sampling Plans for Current Research	RDD Telephone Sample	Address-Based Sample	Traditional Internet Probability Sample	GfK Knowledge Panel® Probability Sample	Ipsos Nonprobability Sample	NPD Nonprobability Sample
General population coverage	Medium	High	Low	Medium/ High	Medium	Medium
Known probability of selection	Medium	High	Low	High	Low	Low
Response rate and cooperation rate	Low	Medium	UNK	Low	UNK	UNK
Cost per complete	High	Medium	Low	Medium	Low/ Medium	Medium

Survey Mode

As is the case for sampling source, the choice of survey mode can impact the representativeness of the sample by influencing the demographics of those who choose to actually take or complete the survey. In addition, the burden that a particular choice of survey mode places on respondents can influence the accuracy of the data obtained from the survey for a given respondent. Issues of selection bias and measurement error thus have to be considered when choosing the survey mode.

Web-based questionnaire: Some of the many benefits to online surveys include more rapid and reliable transmission of completed questionnaires as well as more flexibility in skip patterns. This can also reduce respondent fatigue by withholding non-applicable items (Crossley & Winter, 2012; Dillman et al., 2009b). Related to this is the lower cost of administering a web survey versus other modes due to the lack of need to send or code a physical questionnaire or have an interviewer make contact with the respondent. The accuracy of responses may also be improved relative to in-person or telephone surveys because of the ability of respondents to retrieve relevant information—a benefit that results from the ability of respondents to answer web-based questions when convenient. Another benefit of web-based (and mail-based) surveys is that social desirability effects that result from the presence of an interviewer can be mitigated. On the other hand, interviewers can diminish the effects of respondent confusion by helping clarify ambiguous questions or following up on inconsistent responses—advantages that a web-based or mail based survey may lack. There is also evidence of a "primacy" effect in responses to visual based surveys (i.e., web- and mail-based ones), where respondents are more likely to pick the first option given in a list of discrete responses (Dillman et al., 2009).

Mail-based surveys: Decreased coverage of telephones and difficulty in estimating coverage in web panels has led to increased use of address-based sampling (ABS) and, subsequently, mail-based surveys. In addition to this positive association with ABS, mail surveys have actually maintained relatively higher response rates than telephone- and web-based surveys. However, mail-based surveys also have a number of weaknesses. They are generally less flexible when it comes to skip logic than web-based surveys. In addition, a mail-based survey may be significantly more costly than a web-based one in terms of time and money because of the significant variable costs needed to publish survey and mail pieces, transport these pieces between survey administrators and respondents, and code the responses. This disparity is likely to be greatest for larger survey efforts requiring big sample sizes and/or significant follow-up.

In-person surveys: Although there may be many variations of this approach, for the current project in-person surveys may involve interviewers waiting outside of restaurants and other service-industry establishments with the purpose of asking patrons a battery of items associated with their tipping behavior. This approach could allow for immediate recall of a behavior as well as confidence ratings of the data if the interviewer was trained (as done by the Bureau of Labor Statistics [BLS] in "the New Orleans Test"), thus possibly ensuring more reliable data (Crossley & Winter, 2012). However, the cost can be quite prohibitive, particularly with respect to the number or survey administrators and/or transportation required to ensure that different demographic groups, establishment types, and geographic areas are properly represented in the sample. BLS, in particular, has conducted a number of longitudinal in-person studies over the years under its National Longitudinal Surveys program, and has used a number of techniques to keep respondent attrition low. These techniques include giving their researchers access to local resources to track down any respondents who might have moved or passed away since the previous survey and corresponding with the participant to encourage survey compliance (thank-you letters and pamphlets highlighting the data and knowledge gleaned from the survey effort are two examples). To mitigate the social desirability issues previously discussed with this method, BLS has incorporated computer-based response options so respondents can listen to sensitive questions with headphones and type in their responses without their interviewer's knowledge.

Phone survey: Phone surveys can either be administered by working off a purchased consumer directory or through RDD. With the advent of cell phones, many households no longer use landline phones, and that makes a portion of the population difficult to reach in a cost-effective manner (Pew, 2012).[4] RDD also lends itself to difficulty in measuring non-response bias given the lack of knowledge of the sample frame and, specifically, the nonrespondents. Overcoming low response rates (and potential selection bias) can require frequent calls, increasing the cost of this mode. Another potential issue is that these types of real-time surveys do not give respondents enough time to refer to their schedules or other sources of information concerning past expenditures compared with web and mail surveys (Crossley & Winter, 2012). This will tend to undermine data quality.

Diary study: Following Pearl and McCrohan (1984) and McCrohan and Pearl (1991), a diary panel can be used to provide data over a certain time span for each observation (i.e., tipping behavior of interest) and has been used for both servers and customers in the past. The fact that respondents are expected to record their expenditures near the time when the expenditure was made can mitigate the effects of recall on response accuracy that would plague a recall-based survey. However, this lack of recall bias can come at the cost of not properly capturing seasonal fluctuations in expenditure and tipping behavior if the diary period is short and/or infrequent. In addition, research burden on the participant is quite high and compliance (in the form of attrition and recorded expenditures) significantly drops off over time. It is also possible that the act of recording

[4] Pew Research (2012) "Assessing the Representativeness of Public Opinion Surveys." http://www.people-press.org/2012/05/15/assessing-the-representativeness-of-public-opinion-surveys/

expenditures may induce a downward trend in expenditures over time (Crossley & Winter, 2012). This learning effect is a well-known research confound whereby subjects modify their future behavior in response to the knowledge and skill they gain by being part of the study.

Mixed-mode surveys: Using multiple survey modes has the potential benefit of increasing response rates because of differences in mode preferences across different respondents (Dillman et al., 2009; Crossley & Winter, 2012). However, given that mode has an effect on response quality, data gathered using different modes will not necessarily produce comparable responses (Dillman et al., 2009). Measurement error due to mode effects may be exacerbated if modes with low degrees of recording error are combined with a mode with a high degree of recording error versus the use of a single mode with a low degree of recording error. There is consequently a trade-off between nonresponse/selection and the potential for measurement error.

Table 3. Summary of Trade-Offs for Alternative Consumer Study Modes

Qualities of Modes for Current Research	Web-Based Questionnaire	Mail-Based Paper Survey	In-Person Interview	Phone Survey	Diary Study*
Interviewer effects	Low	Low	High	High	Low
Learning/Testing effects	Low	Low	Low	Low	Medium
Respondent controls when to participate (at a convenient time)	High	High	Low	Low	Medium
Dynamic question branching	High	Low	High	High	High
Quick data turnaround	High	Low	Low	High	Medium
Immediacy of recall	Low	Low	High	Low	High
Administration costs	Low	Medium	High	Medium	High

*Through electronic diary

Study Design

To obtain a picture of expenditure and tipping behavior that is representative of a given period of interest, several study designs can be employed. These designs would differ with respect to the number of times individual respondents are interviewed, the period over which the interviews take place, and the length of the period over which the respondent is required to recall their tipping behavior. A longitudinal, or diary, study would involve surveying individual respondents about their tipping behavior multiple times over the course of the period of interest. A cross-sectional study involves surveying each respondent once over a short period, while requiring that they recall their tipping behavior for the entire period of interest. Finally, a repeated cross-section would only require that respondents provide information about their tipping behavior for the period immediately preceding the interview, but the interviews would be conducted over the entire period of interest.

Longitudinal: A longitudinal study requires repeated observations of the same subject over a specific length of time. Because the same subjects are tracked over time in a longitudinal study, researchers can more reliably attribute a change in behavior to an observed variable. In terms of the proposed methodologies, a longitudinal diary study could illuminate changes over the course of a week or across seasons in consumers' tipping behavior. Longitudinal studies could also be used to track the tipping rate over time with a multiyear effort. In addition, when examining the causes of tipping behavior, longitudinal data allows one to control for unobserved individual level factors that affect tipping, enhancing the researchers' ability to make causal inferences. However, these two latter benefits may not be relevant for the purposes of this project. Asking participants to record their tipping behavior for every service-related purchase immediately afterward over a specified period of time (e.g., one week) would allow for data collection among several different service industries without the need for recall. However, attrition among longitudinal studies is certainly higher and places a higher burden on the respondent. Furthermore, longitudinal studies tend to be more expensive than cross-sectional studies that merely ask for participation for a short duration of time.

<u>Cross-sectional</u>: Unlike longitudinal studies, cross-sectional studies do not utilize repeated observations of the same respondent. Instead, cross-sectional studies aim to survey people of different populations at one point in time, allowing for researchers to compare different populations simultaneously. At another time, the researcher surveys a different sample that is estimated to be congruent to the previously surveyed sample. This form of surveying avoids the high costs and high attrition rates associated with longitudinal studies. All of the proposed data collection methods could potentially use a cross-sectional approach. Mail-, online-, and phone-based surveys frequently use single contacts with participants in order to aggregate data for a given population. Similarly, diary studies can take a cross-sectional approach in the sense that participants are asked to provide feedback about tipping behavior over a 24-hour period. In the process, they would rely less on respondent recall, but avoid the burden of high costs and attrition associated with a longitudinal diary study. However, it is more difficult to be sure that changes in variables of interest within populations are due to outside factors, because respondents are being grouped as opposed to following the same respondents over time. In addition, estimates derived from a single cross-sectional survey with a short recall would not accurately reflect annual tipping rates if expenditure and tipping rates vary by season or day of the week.

<u>Repeated Cross-Sectional</u>: Repeated cross-sectional studies, also known as synthetic panels, offer an alternative to longitudinal and single cross-sectional studies (see Parker, Souleles, & Carroll (2012). Data from multiple cross-sections of survey data would be pooled and respondents sorted into strata defined by multiple, unchanging characteristics (gender, ethnicity for individuals, establishment type and location for establishments/managers). Changes in mean outcome variables (bill size/tipping) for individual strata could then be tracked over time to discern seasonal trends in reported tipping. Unlike single cross-sectional studies, this design/methodology allows variation over time in respondents' tipping (in the case of a consumer) or tip reporting (in the case of server/establishment surveys). In addition, these types of studies are less susceptible to issues associated with longitudinal studies related to survey nonresponse and attrition. The original tipping studies conducted by IRS/NPD, while using data collected through a diary, treated their data as a repeated cross-section for the purpose of analysis.

Analytic Considerations

The goal of the IRS tipping project is to produce estimates of establishment and/or employee tip income that will inform the development of policies that encourage tip reporting. Given that tip income, tip reporting propensity, and optimal policies to encourage tip reporting are likely to vary by sector and geography, estimates of tip income at the industry-location level will likely be more useful to the IRS than more aggregated data. As individual establishments and employees may be less likely to provide accurate responses to surveys that ask about tip income (Simpson, 1997), consumers have been the focus of past research in this area. However, because compliance-based policies are inevitably going to focus on specific types of establishments and locations, consumer tipping data is only useful if consumer tipping can be linked to particular industry-locations. Given that most establishments likely draw the bulk of their customers and tipping revenue locally, this implies that to produce accurate estimates of tipping revenue for particular industry-locations, estimates will have to be produced for relatively small geographic units. This section considers two methods of estimating tipping rates for small geographic areas and their implications for the design of the data collection instrument: Disaggregated Means (DM), and Multilevel Regression and Post-Stratification (MRP).

Disaggregated Means: The simplest approach to estimating tipping rates for particular geographic units would be to simply take the mean tipping rate for all respondents located in a particular geography. Specifically, the estimate is calculated as:

$$\mathrm{T_{jk}} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{T_{ijk}}$$

Where $\mathrm{T_{ijk}}$ is the tipping rate of individual *i* for sector *j* in location *k* and *n* is the number of individuals in location *k*. Besides its simplicity, the advantage of DM is that it makes few assumptions relative to a model-based approach such as MRP (see below). The disadvantage of DM is that the number of observed tipping incidents for a given establishment type/location strata may be very small given a nationally representative survey of typical size (N = 5000).[5] Consequently, bill sizes/tip rates for given sectors/locations from the survey will likely be particularly noisy for a nationally representative sample. Indeed, for very small levels of geographic aggregation, such as counties, there may be no observations for a given establishment type to make the estimate. For this reason, a model-based strategy, like that undertaken by McCrohan and Pearl (1991), may, under certain assumptions, be used to extract precise predictions of tipping rates at a more disaggregated level. One such modeling-based approach, MRP, is discussed below.

Multilevel Regression and Post-Stratification: One means of linking customer-level tipping data to establishments while mitigating issues related to noise in small strata is MRP (Gelman & Little, 1997[6]; see Buttice and Highton, 2013, for a recent review and critique). MRP has attained popularity

[5] Buttice, M. K., & Highton, B. (2013). How does multilevel regression and poststratification perform with conventional national surveys? Forthcoming, *Political Analysis*.

[6] Gelman, A., & Little, T. C. (1997). Poststratification into many categories using hierarchical logistic regression. *Survey Methodology, 23* (2): 127-135.

by social scientists who wish to obtain geographically disaggregated estimates of a quantity of interest.

Analyzing consumer tipping data using MRP would first involve estimating models of consumer expenditure and tipping that take the form:

$$\hat{E}_{ijk} = \beta X_{ijk} + \alpha G_k + C_j$$

$$\hat{T}_{ijk} = \beta X_{ijk} + \alpha G_k + C_j$$

Where *E* is the amount spent by respondent *i* for a service in sector *j* in location *k*; T is a tip rate calculated by dividing a reported dollar amount in tips by *E* or by directly asking the respondent for a tip rate; *X* is a set of observable respondent -level demographic variables such as race, socioeconomic status, etc. , that are likely to influence both tipping and expenditure; and *G* is a set of location-specific factors such as whether the location is part of a rural or urban region that capture variability in expenditure and tipping by sector that is not explained by differences between locations in *X*. Locations are defined as the market area of the establishment. Although it is likely that the size of a given market area will vary by establishment, it might be more practical to assume that an establishment draws most of its customers from the county or metropolitan area in which it is located. Finally, *C* is a constant. After estimating model parameters β, α, and *C*, predictions are generated for strata defined by all *N* combinations of values of *X* and *G* covariates. Poststratification is then used to generate an average tipping rate for a given establishment type/ location:

$$T_{jk} = \sum_{s}^{N} \frac{E_{sj} P_{sk}}{\sum_{s}^{N} E_{sj} P_{sk}} T_{sj}$$

Where *P* is the population of a given stratum in a given location, taken from census data. Estimates for the average tipping rate for a given sector/ location is thus the average tipping rate across all strata, weighted by the strata's expenditure at a given establishment type and the proportion of a location's population in the str ata. The benefit of using a quasi linear, additive model to produce predictions for individual strata rather than using nonparametric estimates from the survey is that, if the linear model provides reasonably accurate estimates of expenditure and tipping rates, the resulting strata -level predictions are likely to suffer less from sampling variability in small to moderate sample sizes than nonparametric estimates. The resulting estimated sector -location tipping rates can be multiplied by an establishment 's reported bill size to arrive at a prediction for its tip income. This estimate can then be compared with reported tip income to arrive at estimates of tip reporting.

Note that the model described above is more flexible than that presented in McCrohan and Pearl (1991) insofar as it (1) disaggregates tipping occasions by industry for the purpose of the regression and (2) incorporates consumer -level demographic data into predictions. Although the model in McCrohan and Pearl (1991) only allowed predicted tipping rates to vary by establishment type and by limited degree geography (size of metropolitan area and census division), MRP may produce predictions of tipping rates by establishment type for a location that varies not just by metropolitan status and census d ivision but, because of the poststratification step, also by the demographics of a particular locality.

Integrating Data Collection and Data Analysis

Obtaining usable information from consumer tipping data will require that the design of the data collection instrument anticipate the requirements of the methodology used to analyze the data. With the assumption that this methodology will incorporate features of both a DM and MRP, this section reviews some items to consider when designing a survey instrument.

Observable Variables: The poststratification stage of MRP requires counts of demographic strata defined by the individual-level variables in the regression stage for the geographic units of interest (i.e., market areas of establishments). Given this requirement, a review of available 2010 Census or 5-year American Community Survey (ACS) data would allow for a determination of what strata counts are available. This will, in turn, inform the construction of the survey instrument to ensure that relevant demographic data is obtained from respondents. If, for instance, we could obtain data on number of individuals of a given age-race-income strata by county, we would want to make sure we could obtain data—either from the respondents or the survey frame on age, race, and income — similar to the original IRS tipping study (Pearl & McCrohan, 1984; McCrohan & Pearl, 1991), so as to post-stratify by income group, age, and region using strata counts taken from Census data.

Geographic Variation: MRP accounts for regional variation in outcomes of interest (in our case, tipping), by including region-level variables that are thought to predict that outcome. To model the effect of region-level variables on tipping, we will require that our survey/diary sample be drawn from variable localities. With respect to geography, the academic literature on tipping has generally focused on differences in tipping between individuals located in metropolitan and nonmetropolitan areas. This suggests that our geographic variable should be some indicator of urban status or population density. However, this might pose a problem for estimating a multilevel regression in a nationally representative sample given that the overwhelming majority of the country's population lives in urban areas. Consequently, it would probably be advisable to oversample rural areas. To do this, however, it will be necessary to define our urban-rural typology before fielding the survey/diary. Specifically, we will want to decide on the urbanization categorization. One simple categorization scheme is the Rural-Urban Continuum Codes (RUCC) produced by the U.S. Department of Agriculture[7]. RUCC codes incorporate information on a county's population density as well as its proximity (adjacency) to metropolitan areas. The advantage of the use of adjacency is that it may better reflect the proximity of an individual residing in a county to large numbers of other people than would be the case if only the county's population density were considered. One downside to the RUCC relative to a simple measure such as population density is that it is tied to counties. If we decide to use a geographic unit other than counties, using the RUCC scheme would require some means of assigning a status to the alternative unit, which would be simple if the unit were nested within counties, but less so if counties were nested within the alternative unit or if the borders did not align with counties, such as in the case of Designated Marketing Areas (DMA).

Temporal Variation: If tipping is seasonal as past research has suggested, computing an annual average estimate of tipping would be complicated by the potential unrepresentativeness of the sample with respect to tipping. This would be the case within a short recall cross-sectional survey to differences in propensity to respond across the year to the day of the week, or in a diary panel

[7] http://www.ers.usda.gov/data-products/rural-urban-continuum-codes/documentation.aspx#.UrMWBfRDu6M

because of attrition. Although this may be mitigated by modeling tipping and expenditure behavior using time effects in order to create a synthetic panel (in the case of repeat cross sections), if the lack of variation is extreme enough, then parameter estimates on the time effects will be imprecise. We might thus want to consider stratifying the sample over days of the week and the year (in the case of a repeat cross section) or have some means of mitigating panel attrition, perhaps by oversampling individuals from demographic groups that have a high probability of attriting[8] or else by having some procedure in place to bring on additional panelists.[9]

Establishment Types/ Sectors: One of the goals of the project is to examine variation in tipping rates by industry and establishment type. This implies the use of an establishment typology. The degree to which survey design will be affected by the need for an establishment typology will depend on the type of information we can obtain from respondents. If we can obtain the name of the establishment where a transaction took place, we may possibly be able to classify the establishment after the survey has been completed depending upon our needs. If that is not feasible, however, we will likely need to obtain information on establishment type from the consumers. In that case, we will have to design the survey such that the options for establishment classification are intuitive and, perhaps most important, limited enough so as not to increase respondent burden to such a level as to increase nonresponse, attrition, or otherwise undermine response quality. The original IRS/ NPD diary panel (Pearl & McCrohan, 1984) arguably did a good job of dealing with this trade-off. Individuals were asked to classify establishments into one of six broad categories and then, in a second question, asked to name the type of food served. Consequently, respondents were not confronted with a large typology of establishment types in one list. Defining establishment types and eating occasions by multiple dimensions and then having a separate question for each dimension allows for a detailed typology while minimizing respondent burden. The chosen typology will also have to be meaningful such that the parameters relating the individual and geographic variables to expenditure and tips will be precisely estimated (i.e., not heterogeneous) when estimated for a given type. Also, this taxonomy must be extended to include establishments other than restaurants. It is thus important that we consider how we are likely to obtain information on establishment type, as that will likely inform the degree of trade-off between collecting accurate information and the precision of the categorization. Another consideration trade-off with having a large number of establishment types is the potential lack of variation in terms of expenditure and tipping behavior one will see if the number of individuals who actually used the service is too small. Larger sample sizes may be necessary to obtain at least some variation in spending and tipping for establishment types for which individual patronage is infrequent.

[8] Frankey and Hillygus (2013) found that non-White respondents were more likely to attrit from the American National Election Study.

[9] McCrohan and Pearl (1991), for example, used a panel that was replenished quarterly to match strata population targets.

Recommended Approach

Based on the benefits and drawbacks of the methodologies reviewed in this report, the following section provides recommendations for the IRS in developing estimates of tipping and stiffing rates, tipping income and, ultimately, the gap between actual and reported tip income both at the aggregate level and by sector. Given many of the unanswered methodological questions in the literature, this report recommends a two-stage process whereby a small set of methods tests will be conducted prior to full-scale administration. Specifically, we recommend examining the performance of a web-based, repeated cross-section survey administered to both a probability and non-probability internet-based panel. The choice of a probability or nonprobability web panel could be adjudicated in a validation phase (see below).

Sample Source

As discussed in the earlier section, all the sample sources covered (RDD, ABS, or the traditional Internet based samples) have a variety of strengths and weaknesses pertaining to sample-related bias. Although phone and address-based frames may arguably be more representative of the U.S. population as a whole than Internet-based panels, response rates are generally low and have been declining over time (Pew, 2012; Keeter et al., 2006 [10]; Curtin, 2005 [11]). These low response rates would likely become even more problematic if, as is recommended below, a web-based mode is used to conduct the survey, given the author's experience with low conversion rates of individuals recruited using these methods to a web-based survey. Further, these more traditional methods may become less mandatory as traditional Internet-based sampling sources continue to evolve, minimizing deficiencies of idiosyncratic recruiting methods prevalent with single source opt-in panels. In fact, recent research on "blended" approaches that use multiple online respondent sources have been found to yield results more similar to dual frame RDD. [12] In addition, the GfK Knowledge Panel ® continues to use a probability based sampling strategy where the survey frame is derived from the USPS Delivery Sequence File.

While none of these methods has a clear advantage with respect to sample-related bias, the same cannot be said for issues related to cost. As already discussed, recruiting individuals using RDD or ABS are likely to be very resource-intensive. In the case of the former, it might take many attempts to contact a given individual before receiving a response, resulting in high labor costs. In the case of ABS, the requirement that the request be printed and transported to the potential respondent carries obvious costs, and response times may be slow. By contrast, recruiting a sufficient number of individuals from Internet-based panels will likely be less costly because of the panelists' stated willingness to participate and the ease of scaling given relatively low variable costs. Even in the case of the GfK Knowledge Panel ®, which recruits its panelists using more costly ABS methods, the

---

[10] Keeter, S., Kennedy, C., Dimock, M., Best, J., & Craighill, P. (2006). Gauging the impact of growing nonresponse on estimates from a national RDD telephone survey. *Public Opinion Quarterly, 70*, 759-779.

[11] Curtin, R. (2005). Changes in telephone survey nonresponse over the past quarter century. *Public Opinion Quarterly, 69*, 87-98.

[12] Vidmar, J., Bricker, D., Young, C., Clark, J., Roshwalb, A., & El Dash, N. (2013). *Using non-probability online surveys for exit polling: The case of the 2012 U.S. Presidential Elections*. Paper presented at the 68th annual meeting of the American Association for Public Opinion Research (AAPOR), October 7, 2013.

recruitment costs would be lower than those for phone - or address-based frames because of the low costs associated with cont acting individuals through email.

Consequently, we recommend the use of an Internet -based sample. Further, we recommend pilot testing both probability and non -probability samples in an attempt to validate the quality of the data resulting from samples rec ruited from each source.

Survey Mode

With respect to the survey mode, this report recommends the use of a web -based survey. The primary reasons being minimization of measurement error and relative cost. Because the survey will require individuals to record their expenditures and tips and categorize the types of establishment for at least a day, the amount of information they may *potentially* have to recall and enter is substantial. In fact, the shear amou nt of possible survey branches and associated instruction would make a paper-/ mail-based survey extremely burdensome, increasing the probability of nonresponse, attrition, or otherwise incomplete, inaccurate documentation of tipping occasions, undermining the quality of the data. With respect to in -person and phone -based surveys, data quality issues may arise because of interviewer effects as well as the inability of the respondent to invest time in recalling accurate information about his or her tipping be havior. By contrast, a computer -based interface can make finding the type of establishment and entering tipping expenditures relatively easy, through dynamic branching, instruction, and look -ups.

Another clear advantage of web -based modes is related to cos t. In-person, phone -based, and mail - based surveys all have high variable costs which are likely to be substantial due to the large number of people that will be required to estimate tipping rates on low frequency behaviors like casino gambling. By contrast , web-based modes can be scaled at relatively low cost.

Survey Design

The primary considerations for survey design are the ability of a specific design to obtain information on tipping that is representative across both individual and time as well as the degree to which different designs increase respondent burden, and thus risk nonresponse/ attrition and/ or poor data quality. Given these considerations, this report recommends the use of a repeated cross -sectional design. Given that each individual is only surveyed once, in contrast to a consumer diary (longitudinal design), where an individual is expected to record the details of tipping occasions multiple times, respondent burden, and thus the unrepresentativeness of the final samp le can be considerably improved. The one -shot nature of the cross -sectional design may also mitigate the risk that the survey itself will influence behavior. One of the primary benefits of a longitudinal design, the potential to make inferences about the c auses of individual expenditure behavior, is arguably of limited relevance in this context as the IRS is primarily concerned with estimating tipping and stiffing behavior rather than explaining individual differences related to consumer tipping. Finally, t he costs associated with gaining longer term commitments and incentivizing participation can be considerably higher for longitudinal designs.

With the repeated cross -sectional design, we further recommend a short -recall period to increase the accuracy of r ecall, reduce respondent burden, and consequently minimize the role of measurement error. Shorter recall periods mean that the tipping occasion reported by a given respondent is not representative of their yearly tipping. However, because of seasonal differences in tipping behavior and the frequency of tipping occasions for specific industries, the repeated nature of the survey increases the potential for variation in both the days of the week and season for tipping occasions in the sample. This variation then allows for the further development of period-specific estimates of tipping using poststratification weighting techniques.

To obtain a large enough sample of respondent-day observations to ensure that there is sufficient variability in low frequency tipping occasions for analysis, the number of respondents used in a repeated cross-sectional study may have to be very large or the recall length extended with the implied increase in measurement error. It should be noted that the IRS' initial tipping study conducted 30 years ago roughly averaged 60,000 respondent-day observations each year (approximately 4,200 respondents over a 14 day period each year). Although this sample size was largely driven by the existing NPD diary data collection this IRS study was attached to, this is roughly the magnitude that we would expect would be necessary to adequately capture the "opportunity for tipping" on low frequency behaviors like casino gambling. For example, as seen in Table 4 below, we estimate needing approximately 76,000 respondent-days to capture 350 casino gambling occasions. This would entail 76,000 respondents if the recall was 24 hours and fewer if the recall length was extended. Although we strongly recommend a short recall period, the day or days this represents should be determined in the piloting stage of the study as prior research does not provide explicit guidance on this key detail. Table 4 provides estimates for the sample size required to obtain different frequencies of tipping occasions by sector. A one-day recall is assumed to remain conservative with the projected estimates.

An alternative to relying on a large nationally representative sample to capture sufficient variation in infrequent activities is to oversample from regions where the activity is expected to be more frequent. This strategy would be most suitable for activities like gambling, where establishments are geographically clustered. Potential complications that result from oversampling arise from the fact that individuals residing in gambling localities may not be representative of the total U.S. population with respect to tipping rates. These differences may reflect the fact that gamblers in high-gambling localities are less likely to be on vacation when they gamble. There may also be systematic differences with respect to demographic characteristics between high gambling and low gambling regions that influence gambling-related tipping. In a model-based approach such as MRP, this could be accounted for by including an indicator for residence in a tipping locality as well as an indicator if the individual were on vacation when the gambling took place. If the assumptions of the model were accurate, relevant differences between gamblers in high gambling regions, gamblers in low gambling regions, and those who gamble on vacation could be accounted for in the final estimate through post-stratification. An alternative approach that avoids the model based assumptions would be to calculate a weighted mean tipping rate, where respondents from oversampled localities would be given a smaller weight such that the weighted sample is representative of the national population with respect to geography. However, this would result in a smaller effective sample of gamblers and gambling occasions, which would increase variance in the final estimate, potentially limiting the benefits of oversampling.

Table 4. Estimated Annual Occurrence

				Estimated Number of Occasions for a given Sample Size				
	Occasions per year	Likelihood per day	Required sample for 350	10,000	30,000	60,000	120,000	240,000
Take-out/ fast food*	115.0	0.315	1,111	3,151	9,452	18,904	37,808	75,616
Eating out/ sit down*	68.0	0.186	1,879	1,863	5,589	11,178	22,356	44,712
Salon**	6.3	0.017	20,373	172	515	1,031	2,062	4,123
Hotels/ motels**	0.6	0.002	223,826	16	47	94	188	375
Taxi/ Limo**	0.6	0.002	210,415	17	50	100	200	399
Casino***	1.7	0.005	76,314	46	138	275	550	1101

Notes: * Estimates of o ccasions per day taken from Pearl and Mc Crohan (1984). ** Estimates of o ccasions per day generated from the detailed monthly expenditure file of the Consumer Expenditure Survey [13]. ***Estimate is an average based on data taken from Shinogle, Norris, Park, Volberg, Haynes, & Stokan (2011) and Volberg, Nysse -Carris, and Gerstein (2006) [14].

Next Steps

This report lays out a general recommended approach; it also leaves open a number of key choices — such as the use of a probability or nonprobability sample, the period of recording/ recall, and the type of model (MRP versus DM). These choices are critical as they ma y lead to invalid predictions due to the data (e.g., selection bias and measurement error) and issues with the model (e.g., included variables and functional form assumptions). Both issues can be relatively difficult to remedy after data has been collected . If the data is measured with error or if there is substantial response bias, it will be unclear what precisely is being modeled and additional rounds of data collection might be prohibitively expensive.

If the dataset does not contain a large range of po tentially observable respondent characteristics, then testing alternative model specifications might be impossible. For this reason, before settling upon a final method, we believe it will be imp ortant to conduct a set of method studies to examine the validity and feasibi lity of our recommended approaches .

[13] For the purposes of calculating the number of occasions per year, a non -zero monthly expenditure on a given activity is assumed to equate with one occasion in that month for the individual respondent. The number of occasions per year is then the fraction of p erson-months with non -zero ex penditure multiplied by 12. Note the assumption that an individual engages in a maximum one expend iture a month likely depresses the number of occasions. Consequently, these estimates should be viewed as conservative.

[14] Shinogle, J., Norris, D. F., Park, D., Volberg, R., Haynes, D., & Stokan, E. (2011). Gambling prevalence in Maryland: A baseline analysis. Volberg R.A., Nysse -Carris K.L., and Gerstein D.R. (2006). 2006 California Problem Gambling Prevalence Survey. Estimates based on Table 4.15 on pg. 26 and Table 3 on pg. 31 , respectively. Respondents who list "Past Year Participation ," assumed to gamble at a Casino once pe r year; "Monthly Participation," 12 times per year; "Weekly Participation ," 52 times a year. Note that c asino gambling is legal in both Maryland and California. In addition, California is in close proximity to Nevada. Consequently, the fraction of the population who reports gambling at a casino, and especially those who visit the casino frequently, may be la rger than in the national population. As a result, Table 4 may inflate the number of casino gambling occasions that would be obtained in a na tionally representative sample.

Instrument Development: The first step, which would focus on instrument development and choice of recall length, should occur even before a pilot study is initiated. Survey usability testing can be used to identify problems in the self-administration of the surveys and interpretation of survey items and instructions. In its most basic form, usability testing is a pretest in which participants are asked to think aloud while completing the survey instrument and describe their thought process for determining their answer to the survey item. Hearing participants vocalize this "inner speech" provides insight into the respondents' understanding of the question wording, response categories, and survey organization. After completion of the survey, additional cognitive probing can be done to explore understanding of concepts that did not emerge during the "think aloud" process. If issues are identified, the survey can be refined and additional cognitive interviews will be conducted to verify the changes. In this respect, survey development and usability should be performed iteratively.

One of the primary focuses of this test would be to understand the process through which people recall their expenditures in order to make a consistent decision on one or multiple days of recall. If usability testing, for example, demonstrates that user's performance is similar in both one- and two-day recall, we would suggest including this variable in subsequent pilot testing.

Pilot Testing: Once an instrument or instruments have been developed, we would suggest a pilot test to further examine the measurement characteristics of the instrument while also examining the use of probability and nonprobability internet panels. As discussed above, the trade-offs between cost and quality are not entirely clear between these two sample sources and would benefit from an empirical test prior to full scale implementation. In addition, to the degree that the usability testing yields ambiguous results with respect to the effects of recall length on accuracy, recall length may also be used to define the set of instruments subject to testing in the pilot phase. We would recommend conducting a test of approximately 20,000 respondent-days (10,000 each method), within one month, spread over approximately 30 Designated Market Areas (DMA). Initial analyses would include an examination of relative differences in estimates, indicators, as well as response characteristics. Although this will provide some evidence as to the consistency of these methods, it will provide little by way of validation evidence. For this, it will be critical to identify a benchmark data source.

One potential source of validation data is point of sale (POS) electronic billing records. Organizations like Restaurant Sciences collect electronic billing records/guest checks to compile useful data for the restaurant industry. This data, including bill and tip totals, can also be purchased by third parties. However, because not all tips are paid using a credit or debit card, such estimates will likely provide an underestimate of total tip income, and therefore cannot be taken as accurate. One way of generating comparable predictions would be to only model expenditures and tipping rates that are paid using a debit or credit card. The dependent variables of the tipping and expenditure models would then be zero if the payment or tip were made using cash, and equal to the amount expended or the tip rate otherwise.

One issue with this type of validation is that the validation metrics would only apply to electronic tipping in restaurants, and would not necessarily say much about the ability of the model to predict nonelectronic, nonrestaurant tip revenue. This type of selection bias would be expected if restaurants that report electronic payments were systematically different from those that do not, with respect to their tip rates. This would be the case, for example, if restaurants with the means to report their electroni c tips were generally better organized. Better organization may be reflected in better service quality and thus higher tips. Another issue with this type of validation is that electronic payment data will likely only be available for restaurants, and thus this data has less to say about the validity for model predictions for nonrestaurant sectors. With these caveats in mind, this out -of-sample data source could provide an extremely valuable source of validation independent of respondent survey data.

Model Validation: Implementing a MRP approach places additional requirements on the data collection instrument. Specifically, for the model to be estimated, the sample will likely have to be stratified geographically in order to obtain variation in the geographic variables. This takes the unweighted sample away from being representative and thus potentially leads to less precise national-level estimates. In addition, depending on the proposed model specification, obtaining information for the individual or geograp hic variables may increase respondent burden and thus the risk of non-response or attrition. Consequently, model based approaches, and specifically MRP, should be validated in the Pilot stage with respect to its ability to predict regional -level tipping ra tes.

In the spirit of Buttice and Highton (2013), a potential means of validating the model would be to use the disaggregate mean estimates of tipping from relatively large Restaurant Sciences samples for a set of approximately 30 geographic regions. The number of observations in the given region will be larger than in the primary survey, allowing more precise, non -parametric estimates of tipping behavior in that region. Regions should be chosen for the validation exercise based on dimensions relevant to tipping rates. Specifically, based on prior literature on tipping, we may believe population density or proximity to an urban center is associated with tipping rates. In that case, the sample of validation regions should vary with respect to their level of urbanization. Note that, because the limited number of observations in the pilot sample, urbanization categories may have to be more aggregated than for the final sample in order to obtain sufficient variation in the geographic covariates (i.e. to obtain observations from less dense, rural regions). If the additive assumptions underlying MRP hold, the MR estimates would be expected to look similar to estimates from these region-specific surveys. Of course, this latter validation step does not account for p otential systematic measurement error that can affect the accuracy of responses to any survey.

The deviation between the prediction and the 'observed' establishment level revenue can be modeled using establishment -level and locality level covariates to pr ovide further guidance with respect to sources of bias. Specifically, we can estimate:

$$|T_o - \overline{T_{jk}}| = \beta O_{ojk} + \alpha G_k$$

In this equation T_o, is the observed tip rate of restaurant o. The left hand side is therefore the difference between the predicted tip rate of est ablishments in its sector and locality. We model this as a function of both establishment -specific characteristics, O, and locality characteristics (G). Note that the locality definitions and characteristics do not have to match those in the models of consumer tipping behavior. This is important because it allows us to incorporate additional geographic information that explains model error. We might find, for instance, that zip -code tabulation area income explains some of the error in the predicted tip rate s. In that case, that would suggest in the full survey, we would want to ensure that we are able to identify the zip code of the respondents for the purpose of modeling. We might also find that, within establishment types, organizational features such as the size of the establishment affects error. To account for this, for the final data collection instrument, we might want to ensure that we are able to collect relevant information about the establishment in order to incorporate those characteristics into our sector typology for the purposes of either DM or MRP, even if it comes at the price of increased respondent burden and risk of selection bias.

Appendix A – Reviewed Articles

Alm, J., & Embaye, A. (2013). Using dynamic panel methods to estimate shadow economies around the world, 1984–2006. *Public Finance Review, 41* (5), 510–543. Themes: METHODOLOGY

Alm, J., & Erard, B. (2013). *Using public information to estimate informal supplier income*. Working paper. Themes: METHODOLOGY, INDUSTRY/SERVICE

Alm, J., & Jacobson, S. (2007). Using laboratory experiments in public economics. *National Tax Journal*, *60*(1), 129–152. Themes: METHODOLOGY

Anderson, J. E., & Bodvarsson, O. B. (2005). Do higher tipped minimum wages boost server pay? *Applied Economics Letters, 12*, 391–393. Themes: GEOGRAPHY, NATIONAL AVERAGE TIPPING RATES

Anderson, J. E., & Bodvarsson, O. B. (2005). Tax evasion on gratuities. *Public Finance Review, 33*, 466–487. Themes: GEOGRAPHY

Ayres, I., Vars, F. E., & Zakariya, N. (2005). To insure prejudice: Racial disparities in taxicab tipping. *The Yale Law Journal, 114*, 1613–1674. Themes: RACE/ETHNICITY

Azar, O. H. (2007). The social norm of tipping: A review. *Journal of Applied Social Psychology*, *37*(2), 380–402. Themes: BILL SIZE

Bodvarsson, O. B., & Gibson, W. A. (1997). Economics and restaurant gratuities: Determining tip rates. *American Journal of Economics and Sociology*, *56*(2), 187–203. Themes: INDUSTRY/SERVICE, BILL SIZE, GEOGRAPHY

Borzekowski, R., & Kiser, E. K. (2008). The choice at the checkout: Quantifying demand across payment instruments. *International Journal of Industrial Organization*, *26*(4), 889–902. Themes: GEOGRAPHY

Boyes, W. J., Mounts, W. S., Jr., & Sowell, C. (2004). Restaurant tipping: Free-riding, social acceptance, and gender differences. *Journal of Applied Social Psychology*, *34*(12), 2616–2625. Themes: INDUSTRY/SERVICE, GENDER, INCOME

Brewster, Z. W. (2012). Racialized customer service in restaurants: A quantitative assessment of the statistical discrimination explanatory framework. *Sociological Inquiry*, *82*(1), 3–28. Themes: RACE/ETHNICITY

Brewster, Z. W., & Mallinson, C. (2009). Racial differences in restaurant tipping: A labour process perspective. *The Service Industries Journal, 29* (8), 1053–1075. Themes: RACE/ETHNICITY

Chapman, G. B., & Winquist, J. R. (1998). The magnitude effect: Temporal discount rates and restaurant tips. *Psychonomic Bulletin & Review, 5* (1), 119–123. Themes: BILL SIZE, INDUSTRY/SERVICE

Crossley, T. F., & Winter, J. K. (2012). Asking households about expenditures: What have we learned? In *Improving the measurement of consumer expenditures*, National Bureau of Economic Research. Themes: METHODOLOGY

Curtin, R. (2005). Changes in telephone survey nonresponse over the past quarter century. *Public Opinion Quarterly,* 69, 87-98.

Davis, S. F., Schrader, B., Richardson, T. R., Kring, J. P., & Kieffer, J. C. (1998). Restaurant servers influence tipping behavior. *Psychological Reports, 83*, 223–226. THEMES: GENDER, GEOGRAPHY

Even, W. E., & Macpherson, D. A. (in press). The effect of the tipped minimum wage on employees in the U.S. restaurant industry. *Southern Economic Journal*. Themes: GEOGRAPHY, NATIONAL AVERAGE TIPPING RATES, METHODOLOGY

Fan, W., & Yan, Z. (2010). Factors affecting response rates of the web survey: A systematic review. *Computers in Human Behavior*, *26*, 132–139. Themes: METHODOLOGY

Feinberg, R. A. (1986). Credit cards as spending facilitating stimuli: A conditioning interpretation. *Journal of Consumer Research, 13* (3), 348–356. Themes: CASH VERSUS CREDIT

Fernandez, G. A. (2004). The tipping point —gratuities, culture, and politics. *Cornell Hospitality Quarterly, 45* (1), 48–51. Themes: TIPPING KNOWLEDGE, RACE/ ETHNICITY

Filion, K., & Allegretto, S. A. (2011). *Waiting for change: The $2.13 Federal subminimum wage* (Briefing Paper No. 297). Economic Policy Institute and Center on Wage and Employment Dynamics. Themes: GEOGRAPHY, NATIONAL AVERAGE TIPPING RATES, GENDER

Frankel, L. L., & Hillygus, D. S. (2013). Looking beyond demographics: Panel attrition in the ANES and GSS. *Political Analysis*. Advance online publication. doi:10.1093/ pan/ mpt020 Themes: METHODOLOGY

Frash, R. E., Jr. (2012). Eat, drink, and tip: Exploring economic opportunities for full-service restaurants. *Journal of Foodservice Business Research, 15*, 176–194. Themes: INDUSTRY/ SERVICE

Garrity, K., & Degelman, D. (1990). Effect of server introduction on restaurant tipping. *Journal of Applied Social Psychology, 20* (2), 168–172. Themes: CASH VERSUS CREDIT

Green, L., Myerson, J ., & Schneider, R. (2003). Is there a magnitude effect in tipping? *Psychonomic Bulletin & Review, 10* (2), 381–386. Themes: INDUSTRY/ SERVICE, BILL SIZE

Greenberg, A. E. (2014). On the complementarity of prosocial norms: The case of restaurant tipping during the holidays. *Journal of Economic Behavior & Organization, 97*, 103–112. Themes: GEOGRAPHY

Harrison, G. W., & List, J. A. (2004). Field experiments. *Journal of Economic Literature*, *42*(4), 1009–1055. Themes: METHODOLOGY

Hill, D. J., & King, M. F. (1993). An exploratory investigation into consumer knowledge of tipping etiquette: Accuracy, antecedents and consequences. In W. Darden & R. Lusch (Eds.), *Proceedings of the symposium on patronage behavior and retail strategy: Cutting edge III* (pp. 121–135). Themes: TIPPING KNOWLEDGE

Jargon, J. (2013, September 4). IRS rule leads restaurants to rethink automatic tips. *The Wall Street Journal*. Retrieved from http://online.wsj.com/news/articles/SB10001424127887323893004579055224175110910. Themes: SERVICE CHARGE

Keeter, S., Kennedy, C., Dimock, M., Best, J. & Craighill, P. (2006). Gauging the impact of growing nonresponse on estimates from a national RDD telephone survey. Public Opinion Quarterly, 70, 759-779.

Kerr, P. M., Domazlicky, B. R., Kerr, A. P., & Knittel, J. R. (2006). An objective measure of service and its effect on tipping. *The Journal of Economics*, *32*(2), 61–69. Themes: INCOME, GENDER, CASH VERSUS CREDIT

Klee, E. (2004). *How people pay: Evidence from grocery store data*. Federal Reserve Board. Retrieved from http://www.newyorkfed.org/research/conference/2006/Econ_Payments/Klee_b.pdf Themes: AGE, INCOME

Kleven, H. J., Knudsen, M. B., Kreiner, C. T., Pedersen, S., & Saez, E. (2011). Unwilling or unable to cheat? Evidence from a tax audit experiment in Denmark. *Econometrica*, *79*(3), 651–692. Themes: METHODOLOGY

Koku, P. S. (2005). Is there a difference in tipping in restaurant versus non-restaurant service encounters, and do ethnicity and gender matter? *Journal of Services Marketing, 19* (7), 445–452. Themes: INDUSTRY/SERVICE, GENDER, RACE/ETHNICITY

Koku, P. S. (2007). Some significant factors that influence tipping in service encounters outside the restaurant industry in the United States. *Services Marketing Quarterly, 29* (1), 23–45. Themes: INDUSTRY/SERVICE

Lynn, M. (1988). The effects of alcohol consumption on restaurant tipping. *Personality and Social Psychology Bulletin, 14* (1), 87–91. Themes: BILL SIZE, INDUSTRY/SERVICE

Lynn, M. (2004). Black-White differences in tipping of various service providers. *Journal of Applied Social Psychology, 34* (11), 2261–2271. Themes: INDUSTRY/SERVICE, RACE/ETHNICITY

Lynn, M. (2004). Ethnic differences in tipping: A matter of familiarity with tipping norms.

Cornell Hospitality Quarterly, 45 (1), 12–22. Themes: TIPPING KNOWLEDGE, RACE/ETHNICITY

Lynn, M. (2006). Geodemographic differenc es in knowledge about the restaurant tipping norm. *Journal of Applied Social Psychology, 36* (3), 740–750. Themes: TIPPING KNOWLEDGE, RACE/ETHNICITY, AGE, INCOME, EDUCATION, GENDER, GEOGRAPHY

Lynn, M. (2006). Tipping in restaurants and around the globe: An interdisciplinary review. In M. Altman (Ed.), *Handbook of Contemporary Behavioral Economics: Foundations and Developments* (pp. 626–643). M. E. Sharpe Publishers. Themes: CASH VERSUS CREDIT, BILL SIZE, RACE/ETHNICITY, GENDER

Lynn, M. (2011). Race differences in tipping: Testing the role of norm familiarity. *Cornell Hospitality Quarterly, 52* (1), 73–80. Themes: RACE/ETHNICITY

Lynn, M. (2012). The contribution of norm familiarity to race differences in tipping: A replication and extension. *Journal of Hospitality & Tourism Research* . Advance online publication. doi:10.1177/1096348012451463. Themes: RACE/ETHNICITY

Lynn, M. (2013). A comparison of Asians', Hispanics', and Whites' restaurant tipping. *Journal of Applied Social Psychology, 43* (4), 834–839. Themes: BILL SIZE, RACE/ETHNICITY

Lynn, M., & Gregor, R. (2001). Tipping and service: The case of hotel bellmen. *International Journal of Hospitality Management, 20* , 299–303. Themes: INDUSTRY/SERVICE

Lynn, M., & Latane, B. (1984). The psychology of res taurant tipping. *Journal of Applied Social Psychology, 14* (6), 549–561. Themes: CASH VERSUS CREDIT, GENDER, BILL SIZE

Lynn, M., & McCall, M. (2000). Gratitude and gratuity: A meta -analysis of research on the service - tipping relationship. *The Journal of Soci o-Economics, 29*(2), 203–214. Themes: METHODOLOGY

Lynn, M., & McCall, M. (2000). *Beyond gratitude and gratuity: A meta -analytic review of the predictors of restaurant tipping* . Working paper, School of Hotel Administration, Cornell University. Themes: CASH V ERSUS CREDIT, BILL SIZE, GENDER, RACE/ETHNICITY

Lynn, M., & Thomas -Haysbert, C. D. (2003). Ethnic differences in tipping: Evidence, explanations, and implications. *Journal of Applied Social Psychology, 33* (8), 1747–1772. Themes: RACE/ETHNICITY, AGE, INCOME, EDUCATION, GENDER

Lynn, M., & Williams, J. (2012). Black -White differences in beliefs about the U.S. restaurant tipping norm: Moderated by socio -economic status? *International Journal of Hospitality Management, 31* (3), 1033–1035. Themes: RACE/ETHNICITY

Lynn, M., Zinkhan, G., & Harris, J. (1993). Consumer tipping: A cross -country study. *Journal of Consumer Research, 20* , 478–488. Themes: GEOGRAPHY

McCall, M., & Belmont, H. J. (1996). Credit card insignia and restaurant tipping: Evidence for an associative lin k. *Journal of Applied Psychology, 81* (5), 609–613. Themes: CASH VERSUS CREDIT

McCrohan, K. F., & Pearl, R. B. (1983, August). *Tipping practices of American households: Consumer based estimates for 1979* . 1983 Program and Abstracts Joint Statistical Meetings, Toronto, CA. Themes: NATIONAL AVERAGE TIPPING RATES, GEOGRAPHY, INCOME, CASH VERSUS CREDIT

McCrohan, K. F., & Pearl, R. B. (1991). An application of commercial panel data for public policy research: Estimates of tip earnings. *Journal of Economic and Soci al Measurement, 17* , 217–231. Themes: NATIONAL AVERAGE TIPPING RATES, CASH VERSUS CREDIT, INDUSTRY/ SERVICE, GEOGRAPHY

Morran, C. (2013, September 5). Are these the final days of automatic 18% tips at restaurants? *Consumerist*. Retrieved from http:/ / consumeri st.com/ 2013/ 09/ 05/ are -these-the-final-days-of-automatic-18-tips-at-restaurants/ . Themes: SERVICE CHARGE

Neuman, S. (2013, September 5). IRS to count automatic gratuities as wages, not tips. *NPR*. Retrieved from http:/ / www.npr.org/ blogs/ thetwo -way/ 2013/ 09/ 05/ 219290573/ irs -to-count-automatic-gratuities-as-wages-not-tips. Themes: SERVICE CHARGE

Noll, E., & Arnold, S. (2004). Racial differences in tipping: Evidence from the field. *Cornell Hospitality Quarterly, 45* , 23–29. Themes: RACE/ ETHNICITY

Papp, T. G., & Burkhammer, A. L. (2001, March). *An investigation of server posture and gender on restaurant tipping* . Paper presented at the 22nd Annual Industrial Organizational Psychology and Organizational Behavior Graduate Student Conference, Pennsylvania State University. Themes: GENDER

Parker, J. A., Souleles, N. S., & Carroll, C. D. (2012). The benefits of panel data in consumer expenditure surveys. *National Bureau of Economic Research* . Themes: METHODOLOGY, INDUSTRY/ SERV ICE

Paul, P., & Gardyn, R. (2001). The tricky topic of tipping. *American Demographics* , *23*(5), 10–11. Themes: NATIONAL AVERAGE TIPPING RATES, INDUSTRY/ SERVICE, GEOGRAPHY

Pearl, R. B. (1984). *A survey approach to estimating the tipping practices of consumers* . Special report on regression analysis to the Internal Revenue Service under contract TIR -81-21, Survey Research Laboratory, University of Illinois, Champaign, IL. Themes: GEOGRAPHY

Pearl, R. B., & McCrohan, K. F. (1984). Estimates of tip income in eating places, 1982. *Statistics of Income Bulletin, 3* (4), 49–53. Themes: METHODOLOGY, NATIONAL AVERAGE TIPPING RATES, INDUSTRY/ SERVICE

Pearl, R. B., & Sudman, S. (1983). *A survey approach to estimating the tipping practices of consumers.* Final report to the Inte rnal Revenue Service under Contract TIR -81-21, Survey Research Laboratory, University of Illinois. Themes: NATIONAL AVERAGE TIPPING RATES, CREDIT VS CASH, INCOME, GEOGRAPHY, INDUSTRY/ SERVICE

Pearl, R. B., & Vidmar, J. (1988). *Tipping practices of American households in restaurants and other eating places: 1985–86*. Supplementary report to the Internal Revenue Service under Contract TIR 86-279, Survey Research Laboratory, University of Illinois, Champaign, IL. Themes: CASH VERSUS CREDIT, GEOGRAPHY, INCOME, ED UCATION, AGE

Pearl, R. B., & Vidmar, J. (1988). *Tipping practices of American households in restaurants and other eating places: 1985–86*. Supplementary report to the Internal Revenue Service under Contract TIR 86-279, Survey Research Laboratory, University of Illinois, Champaign, IL. Themes: GEOGRAPHY, INDSUTRY/ SERVICE

Pew Research Center. (2012). Assessing the representativeness of public opinion surveys. 1-25. http:/ / www.people -press.org/ 2012/ 05/ 15/ assessing -the-representativeness -of-public-opinion-surveys/

Rind, B. (1996). Effects of beliefs about weather conditions on tipping. *Journal of Applied Social Psychology, 26* (2), 137–147. Themes: INDUSTRY/ SERVICE

Sanchez, A. (2002). The effect of alcohol consumption and patronage frequency on restaurant tipping. *Journal of Foodservice Business Research, 5* (3), 19–36. Themes: RACE/ ETHNICITY, AGE, CASH VERSUS CREDIT, INDUSTRY/ SERVICE

Schwer, R. K., & Daneshvary, R. (2000). Tipping participation and expenditures in beauty salons. *Applied Economics, 32* , 2023–2031. Themes: SERVICE/ INDUSTRY, INCOME, AGE, GENDER

Seiter, J. S., & Weger, H., Jr. (2013). Does a customer by any other name tip the same? The effect of forms of address and customers' age on gratuities given to food servers in the United States. *Journal of Applied Social Psychology* , *43*, 1592–1598. Themes: ME THODOLOGY, AGE

Simpson, H. (1997). Tips and excluded workers: The New Orleans test. *Compensation and Working Conditions*, Bureau of Labor Statistics, 32 –36. Themes: SERVICE/ INDUSTRY, GEOGRAPHY

Speer, T. (1997). The give and take of tipping. *American Demogra phics, 19* (2), 51–54. Themes: INDUSTRY/ SERVICE, GEOGRAPHY, INCOME, GENDER.

Star, N. (1988). *The international guide to tipping: When, where, and how much to tip in the U.S. and around the world* . New York, NY: Berkley Books. Themes: INDUSTRY/ SERVICE, GEOGRAP HY

Thomas-Haysbert, C. D. (2002). The effects of race, education, and income on tipping behavior. *Journal of Foodservice Business Research, 5* (2), 47–60. Themes: INDUSTRY/ SERVICE, RACE/ ETHNICITY, INCOME, EDUCATION

Appendix B – Annotated Citations

Alm, J., & Embaye, A. (2013). Using dynamic panel methods to estimate shadow economies around the world, 1984–2006. *Public Finance Review, 41* (5), 510–543.

METHODOLOGY: Article uses a model-based approach to estimate the size of the shadow economy for 111 countries across the world for the period 1984 to 2006. The shadow economy is defined as the production of goods and services that are not included in government accounts. To estimate the shadow economy, the authors model the demand for currency, defined as the amount of cash over M2. Cash-based transactions are assumed to be relatively easy to hide from the state. Consequently, economies dominated by shadow activities are expected to also be cash-based, all other things being equal. Cash demand is modeled as a function of proxies for levels of development such as urbanization and per capita income as well as country-level characteristics that are thought to influence the incentive to conceal income from the government (thus increasing the demand cash), including bureaucratic quality, the tax rate, and the level of inflation. The use of panel data provides more observations and thus degrees of freedom than prior country-specific, time-series based analysis of cash demand while also allowing the authors to correct for endogeneity in the predictors. The resulting model is used to predict cash demand as well as a counterfactual set of predictions where there is no incentive to hide income (when government quality, and thus enforcement is at its maximum, the tax rate is zero, and there is no inflation). The predictions for cash demand where there is no shadow economy is subtracted from the total predicted cash demand to arrive at an estimate of cash demand that is due to tax evasion. This estimate is then multiplied by money velocity and divided by GDP to arrive at an estimate of the shadow economy as a fraction of GDP. The results indicate a negative association between the size of the shadow economy and the level of development. However, the mean for OECD (Organization for Economic Cooperation and Development) countries across the entire period is still a substantial 16.9%, and 13.3% for the United States in 2006.

Alm, J., & Erard, B. (2013). *Using public information to estimate informal supplier income*. Working paper.

METHODOLOGY: Authors use responses from the 2001 Current Population Survey (CPS) to estimate informal supplier (self-employment) income and tax noncompliance. Specifically, they develop estimates of national informal supplier income using income information provided by self-employed respondents working in 11 industry categories in which informal suppliers will be prominent. To estimate the income of "Food Caterers and Roadside Stands," the authors use responses from the Bureau of Labor Statistics' Consumer Expenditure Survey (CES). They then compare these estimates with income reported to IRS National Research Program (NRP) in these industries to arrive at industry category-level estimates of tax noncompliance. Supplementary CPS surveys were used to identify second jobs and second job income was imputed based on the assumption that secondary income comprised 26.5% of wages. This fraction was in turn estimated using a subsample of respondents who reported income for both jobs. In addition, self-employed, informal income misclassified as wages was assumed to comprise 4.08% of wages. The resulting estimates of total self-employed income ($156.4 billion) for the 11 CPS industry categories exceeded reported income estimated from the NRP ($50.9 billion), but was lower than an estimate of total income derived from NRP data (reported income + audit detected + estimated non-detected).

Alm, J., & Jacobson, S. (2007). Using laboratory experiments in public economics. *National Tax Journal, 60*(1), 129–152.
METHODOLOGY: Provides a review of literature using laboratory experiments in the field of public economics. The article lays out the requirements for the successful expectation of an experiment studying the effect of incentives on behavior, including control over the experimental environment such that monetary incentives be explicitly linked to behavior, that instructions are clear, that the experiment not be too long or complicated, and that instructions should not use terminology that hints at the research question that the experiment addresses, which the authors argue could potentially influence the subjects' behavior. Common criticisms of experiments include the argument that the mainly university student subject pool of most laboratory experiments is not representative of the wider population whose behavior and motivations the experiment is trying to analyze/explain (though the authors argue this concern is unfounded), that subjects modify their behavior as a result of the awareness that they are participating in an experiment, and that certain factors that affect behavior in the real world, such as the threat of prison time, cannot be plausibly simulated in a laboratory setting. Consequently, results of an experiment may not generalize outside of the laboratory setting. The article also discusses the use of laboratory experiments to address questions related to the determinants of tax compliance behavior. These experiments typically find that audits increase compliance (though there are diminishing marginal returns as the audit rate increases), that the fine rate increases compliance (though the effect is small), and that higher marginal tax rates lead to lower compliance. Higher income is found to lead to greater compliance. Targeted audits have been found to be more effective in increasing compliance than random audits. Democratic participation and an effective social norm supporting tax compliance increase individual compliance.

Anderson, J. E., & Bodvarsson, O. B. (2005). Do higher tipped minimum wages boost server pay? *Applied Economics Letters, 12* , 391–393.

DESIGN OVERVIEW: Authors investigate if there is any difference in server pay between states with varying levels of subminimum wages and tip credits for tipped staff. A probit analysis was used, and there were 100 total observations in the analysis: one observation for waiters and one observation for bartenders for each of the 50 states (Washington, D.C., was not mentioned in the article and was likely excluded). Data was pulled from the Bureau of Labor Statistics "Wages by Area and Occupation" file (additional data was pulled from the U.S. Census Bureau, the National Restaurant Association, and the Bureau of Economic Analysis). Analysis controlled for the percentage of firms exempted from state and federal minimum wage laws, and restaurants' revenue as a proportion of the GDP, in addition to other control variables such as age and whether the state has a state income tax.

AVERAGE TIPPING RATES: OLS regression findings indicate that there was a very small difference between states with no minimum wage or tip credit versus states with no tip credits and wages that exceed federal standards, but that overall there was no noticeable difference between the minimum wage of waiters and reported wages.

Anderson, J. E., & Bodvarsson, O. B. (2005). Tax evasion on gratuities. *Public Finance Review, 33* , 466–487.

DESIGN OVERVIEW: The authors used state-level data from the Bureau of Labor Statistics (BLS) to determine if total reported pay is affected by factors that are hypothesized to affect underreporting of tips. The BLS' Occupational Employment Statistics (OES) surveys are used to estimate the mean and median hourly pay for over 750 occupations, and the authors used restaurant-related occupations for testing their model. Two variables are included to proxy average customer tipping rate (i.e., premium full-service restaurants as a percentage of full restaurants in the state and the percentage of each state's population living in urban areas). They also included several control variables to account for slight differences in job characteristics and locations.

GEOGRAPHY: Reported pay is higher in areas with a higher tipped minimum wage and in states with no income tax. IRS audit rates do not appear to have an effect on reported pay by restaurant employees. The most important result from their analyses was that higher tax rates raise the employee's reported pay, such that one percentage point increase in a state's minimum income tax rate results in servers reporting 13 cents more in pay.

Ayres, I., Vars, F. E., & Zakariya, N. (2005). To insure prejudice: Racial disparities in taxicab tipping. *The Yale Law Journal, 114,* 1613–1674.

DESIGN OVERVIEW : 12 taxicab drivers (6 Black, 4 White, and 2 "other minorities") completed surveys immediately after dropping off customers for a total of 1,066 completed surveys. Tips were calculated by sub tracting the fare from the total cost of the ride. Drivers recorded sex, race, age, passenger dress (proxy for wealth), and driver experience. They also recorded other interaction characteristics, including whether they paid with cash.

RACE/ ETHNICITY : White drivers were tipped 61% more than Black drivers (20.3% versus 12.6%) and 64% more than "other minority" drivers (20.3% versus 12.4%). Black drivers were 80% more likely to be stiffed than White drivers (28.3% versus 15.7%) and "other minority" drivers we re 131% more likely (36.4% versus 15.7%). The mean tipping percentage of Black customers was 42% of the mean tipping percentage of White customers (9.2% versus 21.6%). Hispanic customers' mean tipping percentage was just over half of White customers' mean tipping percentage (12.0% versus 21.6%). Asians tipped 75% of the White customers' mean tipping percentage (16.2% versus 21.6%). White customers stiffed the driver (10.6%) less frequently than Blacks (39.2%), Hispanics (34.3%), and Asians (15.8%). Using a regression analysis and controlling for random driver effects, time, manner, and place effects, Black drivers are tipped 9.1% less than White drivers. In the most complete regression, Black passengers tipped 9% less than White passengers.

Azar, O. H. (200 7). The social norm of tipping: A review. *Journal of Applied Social Psychology, 37* (2), 380–402.

DESIGN OVERVIEW : A literature review of various tipping -related areas, including both theoretical motivations behind tipping behavior and empirical studies on the subject. Areas of focus include the relationship between service quality and tipping behavior, patronage frequency, bill size, service quantity, and other variables.

Bodvarsson, O. B., & Gibson, W. A. (1997). Economics and restaurant gratuities: Deter mining tip rates. *American Journal of Economics and Sociology* , *56*(2), 187–203.

DESIGN OVERVIEW : Authors test several hypothesis derived from economic theory on the determinants of tipping. Data is based on 697 respondents to a survey conducted in 7 Minnes ota restaurants. Data collected included bill and tip size, number of food and drink items ordered, number of people at the table, whether the respondent visited the establishment at least once a month, and an assessment of service quality. To account for potential measurement error in tipping due to social desirability bias, the tip rates reported by customers were passed by the servers who

gave an assessment of their plausibility. Their answer was affirmative. Tip amounts and tip rates were analyzed using both descriptive statistics and multivariate regression analysis.
INDUSTRY/ SERVICE: Tip rates varied across establishments; establishments that were licensed to serve alcohol received higher tips.
BILL SIZE: Tip amount was positively related to tip amount bill size, and only marginally related to service quality, consistent with the existence of a lower bound on the amount customer's tip.
GEOGRAPHY: Tips (amounts and rates) were higher in restaurants located in St. Paul than in St. Cloud, consistent with t ips being higher in larger urban areas.

Borzekowski, R., & Kiser, E. K. (2008). The choice at the checkout: Quantifying demand across payment instruments. *International Journal of Industrial Organization, 26* (4), 889–902.
DESIGN OVERVIEW : Article examinin g roughly 1,500 households over the course of three months from March through May of 2004. The survey was conducted as part of the University of Michigan Survey of Consumers, a telephone -based survey that covers various aspects of consumer behaviors and attitudes. Various scenarios were presented to respondents, including one suggesting that a "flash" debit service has been introduced to see changes in behavior and another that attempts to "age" the cohort to see changes in behavior. Overall, it was reporte d that debit cards were overcoming the use of cash and checks for consumers. However, given that the scenarios presented ask about usage when purchasing items from a supermarket, payment methods will likely be very different for tipping situations, because checks are often not appropriate or accepted for tipping situations or establishments.
GEOGRAPHY: Of the four regions, the West region had the highest predicted market share of debit and credit usage (53% for the two) compared with 46.5% for the South, 41 .6% for the Northeast, and 38.4% in the Midwest.

Boyes, W. J., Mounts, W. S., Jr., & Sowell, C. (2004). Restaurant tipping: Free -riding, social acceptance, and gender differences. *Journal of Applied Social Psychology* , *34*(12), 2616–2625.
DESIGN OVERVIEW : Study investigating tipping behavior using in -person survey intercepts at 18 different restaurant locations, 10 surveys per restaurant. Analysis was used to determine if social acceptance and free -riding influence tipping behavior. Additional variables i ncluded customer gender. In-person intercepts were used at each restaurant, asking respondents various questions about their demographics, the size of their party, whether they are a local resident of the area (used as a proxy to determine if they were a r epeat customer), how often they eat out and how often they have eaten at the restaurant in the past month, and ratings about the quality of their meal.

Respondents were also asked if they had any alcohol or not. Surveys were only asked during dinner hours to maintain consistency; roughly 90% of respondents agreed to respond to the survey, and a third of surveys were confirmed with the servers of the restaurant for accuracy.

Furthermore, restaurants were classified into four different restaurant types

INDUSTRY/SERVICE: Alcohol consumption had a significant impact on the tipping percentage such that respondents who indicated they had consumed alcohol left higher tips.

GENDER: Men tipped less than women, even when other factors were held constant. In addition, men's tips were found to be more significantly influenced by party size.

INCOME: Higher levels of income were related to higher tipping rates.

Brewster, Z. W. (2012). Racialized customer service in restaurants: A quantitative assessment of the statistical discrimination explanatory framework. *Sociological Inquiry, 82* (1), 3–28.

DESIGN OVERVIEW: A paper survey was given to servers from a sample of 18 chain-style restaurants. Overall, 200 completed surveys were gathered. The aim of the survey was to determine whether servers discriminate against customers of various races (based on questions asking if the quality of service will vary by race). The a uthor acknowledged that explicit questions about racial tendencies in this way will lead to some lack of variability in reporting behaviors because people will wish to report in a way consistent with a social -desirability bias. Respondents were given a ser ies of five scenarios (in which the customer race was held constant as Black customers in various configurations) and asked whether the customers were good or bad tippers (on a 5 -point scale). Respondents were also asked what they considered to be good and bad attributes of diners and to provide ratings of the dining behaviors of the Black individuals in the scenarios. Servers were also asked if they preferred to serve various situations (such as groups with or without children, social classes of their clie nts, etc.).

RACE/ETHNICITY: Overall, nearly 1 in 5 servers reported an explicit preference for serving White clients. In addition, on the 4 -point scale regarding service -quality discrimination (1 = never and 4 = always), the mean score was 1.49, indicating that a reasonable number of servers were willing to report some discriminatory behaviors against their customers. Findings seem to indicate that once discriminatory tendencies toward other groups are taken into consideration (such as children, etc.), that servers who report more positivity toward Blacks are less likely to discriminate against them in their service. However, given their use of a proxy variable for discriminatory behaviors, the findings have to be considered with caution.

Brewster, Z. W., & Mallinson, C. (2009). Racial differences in restaurant tipping: A labour process perspective. *The Service Industries Journal*, *29*(8), 1053–1075.
DESIGN OVERVIEW: Literature review of two theoretical frameworks that try to explain the reasons for lower tipping behavior among Blacks. The two frameworks that are discussed are that (1) Blacks are unaware of tipping norms, hence leading to lower tipping behavior and (2) that Blacks tip at lower rates because service providers (i.e., waiters) treat Black customers poorly because they anticipate poor tips, creating a cyclical problem.

Chapman, G. B., & Winquist, J. R. (1998). The magnitude effect: Temporal discount rates and restaurant tips. *Psychonomic Bulletin & Review, 5* (1), 119–123.
DESIGN OVERVIEW: Subjects included 50 undergraduate students participating for course credit. Subjects completed a questionnaire that included two sections: an intertemporal choice and three tipping scenarios. The tipping scenarios comprised a taxi ride, a restaurant dinner, and a haircut. Each scenario included a brief description and asked how much the participant would tip based on bill size. They were presented with four different magnitudes for each tipping setting. Participants were also asked how much they had paid and tipped the last time they had used each of the service scenarios.
INDUSTRY/SERVICE AND BILL SIZE: Tip percentages decreased with bill magnitude for each of the three tipping scenarios, but ANOVA revealed a significant effect of magnitude for the haircut and restaurant dinner scenarios. The magnitude effect (i.e., tip percentages decrease significantly as the bill size increases) was found to be present in both of these scenarios, indicating that participants reported leaving bigger tips for smaller bills.

Crossley, T. F., & Winter, J. K. (2012). Asking households about expenditures: What have we learned? In *Improving the measurement of consumer expenditures*, National Bureau of Economic Research.
METHODOLOGY: Article reviews literature examining the benefits and drawbacks of different methods of collecting household expenditure data through surveys. There is little evidence to suggest the superiority of single survey modes (face-to-face interviews, telephone interviews, self-administered questionnaires); while self-administered questionnaires may increase response rates and quality by allowing respondents time to recall their expenditure patterns and reduce confidentiality relative to modes requiring an immediate response to the interviewer, interviewers may be able to provide more assistance to respondents who have issues with question comprehension. Recall surveys may lead to downward biases in reported expenditure due to poor

recall relative to diaries as well the inclusion of expenditures from before the survey reference period, but diaries may lead to respondent attrition and a decline in the accuracy of responses as time passes due to the greater imposition on respondents. This may lead to a downward bias in expenditure estimates in diaries versus recall surveys, and has been found to be problematic in the case of expenditures on food. Expenditure data collected from diaries with short time frames may also show greater variance due to the fact that respondents report expenditures as they are made, and there may be a large degree of variance in expenditures in short time periods, particularly with respect to infrequent expenditure categories. The keeping of diaries may also influence respondents' expenditure patterns, resulting in biased estimates of population expenditure patterns. Diary respondents may also tend to aggregate different expenditures when they are made at the same time.

The format of survey questions has also been found to have an effect on data quality; open-ended formats lead to rounding of responses, while closed formats may lead respondents to choose categories that they perceive as reflecting their relative expenditures (high spender, high-spending bin) as opposed to their true expenditure. Aggregated expenditure categories tend to lead to lower total expenditure estimates, perhaps due to an inability of respondents to recall every type of expenditure. On the other hand, more disaggregated expenditure categories may put a greater burden on respondents and thus lead to lower quality (less accurate) responses. Using single respondents to solicit information on household expenditures may lead to lower-quality estimates, but using multiple respondents per household may place a greater burden on the household and consequently result in lower response rates. Incentives for completing the survey or diary may increase both response rates and data quality. Data quality can also be improved by asking respondents to reassess their expenditure estimates when they are inconsistent with previously given information, such as total budget.

Davis, S. F., Schrader, B., Richardson, T. R., Kring, J. P., & Kieffer J. C. (1998). Restaurant servers influence tipping behavior. *Psychological Reports, 83*, 223–226.

DESIGN OVERVIEW: Twenty-eight servers from a pair of restaurants (one in a small Midwestern town, 12 servers; and another in an urban area, 16 servers) recorded their tips for a four-week period while alternating whether they stood or squatted by tables in order to determine if that increased tip size. Aside from varying the squat/standing procedure, other descriptive measures including whether the meal was for lunch or dinner and what the gender of the server was were maintained for analysis. Of the 12 servers in the rural area, 7 were women and 5 were men, and there was an even 8/8 split in the urban area. Servers maintained all of the recordings, including the dollar amount of the meal and the tip that they received. Possible issues with this study are that there is no mention of an incentive for the servers to maintain accurate record-keeping and that they might be misreporting their tips as a whole.

GEOGRAPHY: The study found that people from urban areas tipped significantly more than those from rural areas, but because the servers were not able to determine any kind of socioeconomic variables such as income or education, this might be a spurious effect caused by other variables.

GENDER: Female servers received significantly greater tips than male servers (15.6% compared with 14.1%, though this was the smallest of the significant findings).

Even, W. E., & Macpherson, D. A. (in press). The effect of the tipped minimum wage on employees in the U.S. restaurant industry. *Southern Economic Journal*.

DESIGN OVERVIEW: Two sets of regression analyses were run (specifically, the regressions were a version of "difference-in-difference estimation"—additional details and citations about this regression method can be found in the article): one using data from the Quarterly Census of Employment and Wages (QCEW) and the other using data from the Census Bureau's Current Population Survey from 1990 through 2011. The regression equation controlled for changes due to season and various demographic variables that would change earnings in the industry, and accounted for both the federal minimum wage and the subminimum wage, among other factors.

Both data sources have their advantages. The QCEW data is pulled from unemployment insurance reports, ensuring essentially mandatory compliance for reporting. However, this data does not provide work hours for workers, nor does it give characteristics of the workers. CPS data, on the other hand, provides such characteristics, but because of methodology the sample for certain industries and states can be quite small and introduce the possibility of error. Both data sets were acknowledged to have specific strengths and weaknesses for their analysis.

NATIONAL AVERAGE TIPPING RATES: Findings from analyses of both data sources indicate that the salary of tipped workers does increase along with minimum wage increase, though the QCEW data points out that this only occurs among full-service restaurants and is not seen among limited service restaurants. Further findings indicate that increases in the minimum wage for tipped employees has a negative influence on the employment of this population and that raises in this minimum wage lead to reduced hours worked per week in addition to higher wages.

Fan, W., & Yan, Z. (2010). Factors affecting response rates of the web survey: A systematic review. *Computers in Human Behavior, 26*, 132–139.

METHODOLOGY: Article reviews literature addressing factors that affect web response rates. Factors related to survey content include: the sponsor of the survey, with response rates being higher when the survey's sponsor is an academic or government agency; the content of the survey, with surveys asking questions concerning highly salient issues typically receiving higher response rates than those whose subject is less relevant to potential respondents; the length of the survey, with longer surveys having lower response rates. Sample design and contact methods also influence response rates: web panel designs typically yield higher response rates than single-shot surveys, while email-based contact can result in low response rates because of spam filters. However, the use of personalized messages, prenotifications, and reminders can raise response rates. Empirical work examining the influence of incentives (such as an electronically mailed gift certificate) on response rates has generally found small (or even negative) effects on participation. The survey frame also affects response rates, with surveys of the general population generally yielding lower response rates than surveys of specific populations such as employees, though top managers are less likely to respond than lower-level managers/employees. Populations with low socioeconomic status are less likely to respond because of limited Internet access, though this effect persists even after controlling for such access. The personalities of potential respondents also influence response rates, with more conscientious individuals having a greater propensity to respond.

Feinberg, R.A. (1986). Credit cards as spending facilitating stimuli: A conditioning interpretation. *Journal of Consumer Research, 13* (3), 348–356.
DESIGN OVERVIEW: One hundred and thirty-five customers were observed at random intervals over a one-week span at a local restaurant. Servers recorded party size, check amount, mode of payment, and amount of tip. The author also conducted four experiments investigating characteristics of credit card spending, but none of them dealt with tipping or the service industry.
CASH VERSUS CREDIT: A 2 (payment method) x 4 (check size divided into quartiles) ANOVA revealed that when credit card stimuli were present, customers left a significantly higher tip. For each quartile of check size, customers paying with credit cards provided higher tips. Credit card–paying customers, on average, left a tip that was 16.95% of the total bill, while cash-paying customers left a tip that was 14.95% of the total bill.

Fernandez, G. A. (2004). The tipping point—gratuities, culture, and politics. *Cornell Hospitality Quarterly, 45*(1), 48–51.
DESIGN OVERVIEW: Discussion about knowledge of tipping behavior, how the knowledge is passed on, and a discussion about what underlies the racial differences in tipping. Some topics that are

discussed are underlying psychological issues that might be at work within the Black community, including how the segregation of service in restaurants in the past might be the cause of certain behaviors in the present. The author calls for a national study to look at this subject, with enough of a sample to investigate racial differences across different areas with the sufficient detail needed to draw concrete conclusions.

Filion, K., & Allegretto, S. A. (2011). *Waiting for change: The $2.13 Federal subminimum wage* (Briefing Paper No. 297). Economic Policy Institute and Center on Wage and Employment Dynamics.
DESIGN OVERVIEW: Analysis was conducted using the Census Bureau's Current Population Survey from 2008–2009. Descriptive results of reported wages were split by several demographic groups, including worker gender, race, age, education, and across various states with differing levels of wages for tipped employees.
NATIONAL AVERAGE TIPPING RATES: Overall, it was found that states with higher levels of subminimum wages had higher reported hourly wages for waiters and tipped workers than states with lower tipped minimum wages for tipped workers. However, it is worth noting that the median wage of workers was higher in those states overall, indicating that the relative affluence of those states are driving these changes.
GENDER: Demographic splits indicate that while females constitute the majority of tipped workers and waiters (72.9% and 72.4%, respectively) they earn less on average than male workers, particularly among waiters ($9.04 for females and $9.87 for males).

Frankel, L. L., & Hillygus, D. S. (2013). Looking beyond demographics: Panel attrition in the ANES and GSS. *Political Analysis.* Advance online publication. doi:10.1093/pan/mpt020
METHODOLOGY: Article examines the determinants of respondent attrition in the American National Election Studies (ANES), an online panel survey, and the General Social Survey (GSS), a face-to-face interview panel survey using logit regression. Both respondent demographics and survey experience characteristics are included as predictors of attrition. Among the demographic characteristics, age, education, and employment were negatively associated with attrition in the ANES, while non-English preferences and the number of young children were positively associated with attrition. Age and education had a statistically significant negative association in GSS, while foreign born and single member household status were positively associated with the probability of attrition. Among the survey experience characteristics, respondents to the ANES who reported a monetary motivation, had a negative experience, and/or took a long time to complete the survey were more likely to attrite, as were those who refused to answer the survey in the first wave. For the GSS, interviewer

experience was found to be negatively associated with the probability of attrition, and respondents who were interviewed by females were less likely to attrite.

Frash, R. E, Jr. (2012). Eat, drink, and tip: Exploring economic opportunities for full-service restaurants. *Journal of Foodservice Business Research, 15*, 176–194.
DESIGN OVERVIEW: The author pooled point-of-sale (POS) processed guest checks and their associated credit card checks from two restaurants (one fine dining establishment and one casual-theme full-service restaurant). One hundred and fifty checks were randomly selected from each restaurant's weekly pool and each check had to meet several conditions, namely that the checks had to include both food and alcoholic beverages, be from restaurants' dining rooms (i.e., no checks from the bar), be tendered after 5:00 p.m., paid by only one party, and not include any promotional or employee discounting. From the guest and credit card checks, the author recorded reliably accurate information for the guest check dollar amount, percentage of the guest check spent on alcoholic beverage purchases, server's gender, dollar tip amount, and tip percentage. Time the guest check was rendered and day of the week were also recorded.
INDUSTRY/SERVICE: Two hundred and ninety-seven guest checks were included in the final analysis from the two restaurants. The median percentage of the guest check that was attributable to alcoholic beverages was 26.8%, the median guest check was $40.67, and the median tip percentage was 20.6%. A multiple regression was performed to predict the tip percentage from percentage of the guest check used on alcoholic beverages. A positive relationship was found between the percentage of guest check attributable to alcoholic beverages and the tip percentage of the whole guest check.

Garrity, K., & Degelman, D. (1990). Effect of server introduction on restaurant tipping. *Journal of Applied Social Psychology, 20* (2), 168–172.
DESIGN OVERVIEW: Forty-two, 2-person parties that ordered a Sunday brunch at a restaurant were randomly assigned into two interaction conditions. In one condition, the server greets the customer while introducing herself; in the other condition, the server just greets the customer.
CASH VERSUS CREDIT: Customers that used a credit card as a form of payment left, on average, larger tips than those using cash (22.6% versus 15.9%).

Green, L., Myerson, J., & Schneider, R. (2003). Is there a magnitude effect in tipping? *Psychonomic Bulletin & Review, 10* (2), 381–386.

DESIGN OVERVIEW: In order to determine if there is a magnitude effect in tipping (i.e., as bill size increases, percentage tipped decrease s), researchers had two taxicab drivers, four restaurant servers (from two restaurants), and four hair stylists (from two salons) record the total bill size and the amount of the tip for each customer over several months. This amounted to nearly 1,000 serv ice encounters.

INDUSTRY/ SERVICE AND BILL SIZE : The author's regressed percentage tipped on the total amount of the bill for all bills less than $100. The regression slopes were negative in each of the six cases (two taxicabs, two hair salons, and two restaurants), indicating a magnitude effect. Linea r regression results for each of the six establishments demonstrate that as the total bill amounts get even larger, the slope of the regression line becomes less negative, approaching zero.

Greenberg, A. E. (2014). On the complementarity of prosocial norms: The case of restaurant tipping during the holidays. *Journal of Economic Behavior & Organization, 97* , 103–112.

DESIGN OVERVIEW : Data was pulled from all credit card transactions from a res taurant chain in upstate New York over the course of one year. All transactions required both a correct bill and tip amount, so that situations when no tip was left on the credit card were dropped from the analysis (because those situations likely included a cash tip since it was reported that instances of complete "stiffing" among credit card customers were quite rare).

For their analysis, the "holiday period" was determined to be the weeks prior and post -Christmas Day. Furthermore, other holiday days were added into the regression equation as a separate variable. Customers were restricted in the analysis to those who were observed as having dined at least once during the holidays and during the non -holiday period.

GEOGRAPHY : Forthcoming paper looking at wh ether prosocial behaviors (tipping behavior in general and generosity during the holidays) compete with one another, leading to no change in tipping behavior during the holidays, or whether they would complement one another such that people would tip at hi gher rates during the holidays. Overall findings were that people tipped higher during the holidays, but when the population was split, it was determined that this finding was skewed and that while bad tippers tipped better, "good" tippers tipped even more .

Findings were that tips during the holiday period were 3.7% higher than in the non -holiday period (24.3% overall).

Harrison, G. W., & List, J. A. (2004). Field experiments. *Journal of Economic Literature, 42* (4), 1009–1055.

METHODOLOGY: Article discusses the use of field experiments in economic research. In contrast to traditional means of collecting data for the purpose of economic research —such as the use of naturally occurring data, where treatment and control status are not assigned at random, or laboratory experiments, where treatment status is randomly assigned but the setting is artificial — field experiments feature the use of randomly assigned treatment status but in a natural setting. They thus potentially allow the researcher to make causal inferences while simultaneously mitigating issues of external validity that are prevalent in laboratory experiments. The article briefly discusses findings from three types of field experiments that allow for varying degrees of external validity: artifactual field experiments, where the subjects are aware of the experiment and the activity that they undertake does not directly correspond to naturally occurring activities, but where the subject pool represents a naturally occurring population of interest; frame field experiments, where, like artifactual experiments, the subjects are aware that they are participating in an experiment but where the subject's activity in the experiment more closely corresponds to naturally occurring phenomena; and natural field experiments, where the activity induced by the experiment is something the subjects would do naturally and they are simultaneously unaware that they are participating in an experiment, maximizing the chances that observed responses to the treatment would hold outside of the context of the experiment.

Hill, D. J., & King, M. F. (1993). An exploratory investigation into consumer knowledge of tipping etiquette: Accuracy, antecedents and consequences. In W. Darden & R. Lusch (Eds.), *Proceedings of the symposium on patronage behavior and retail strategy: Cutting edge III* (pp. 121–135).
DESIGN OVERVIEW: Sample was roughly 150 business majors ages 20 to 42 used for the analysis. They were asked to provide responses to what the appropriate tipping levels were for various services (not listed by the author, though the articles that they based these "correct" answers on were listed). They created a battery of tipping-related items and used a factor analysis to determine that there were five factors concerning tipping knowledge. Respondents were also asked a series of 27 developed questions on variables that were determined to influence tipping behavior from literature reviews and one-on-one interviews on this subject. The 27 questions were determined to have five useful factors: (1) social tipping orientation (their belief in the "social value of tipping"), (2) tipping experience, (3) tipping confidence (their belief that their knowledge of tipping behavior was correct), (4) tipping response (belief that poor service should receive poor tips, etc.), and (5) parental influence.
TIPPING KNOWLEDGE: Ultimately, most of the factors were not found to be correlated to correct tipping knowledge. The only two that were related were the parental influence (such that those who

learned more from their parents had more correct knowledge) and the age they first tipped (which also makes sense given that the earlier they started tipping, the more guidance they likely got from their parents and practice they have with tipping behavior).

Jargon, J. (2013, September 4). IRS rule leads restaurants to rethink automatic tips. *The Wall Street Journal*. Retrieved from
http://online.wsj.com/news/articles/SB10001424127887323893004579055224175110910.
SERVICE CHARGE: Article reporting on the change in ho w the IRS counts tips automatically added to the bill for large parties and the change that will occur starting in 2014. Under the new rules, restaurants will have to take those automatic tips and add it to the servers' actual wage at the end of the pay cy cle and withhold taxes from it. This means that servers will have to wait for that money, as opposed to getting it at the end of the night, to ensure taxes are filed properly (which could mean less income for servers), and cause more paperwork and costs to restaurants to manage additional records.
This article was later cited by other websites, including NPR and the Consumerist (see Neuman, 2013; and Morran, 2013, citations).

Kerr, P. M., Domazlicky, B. R., Kerr, A. P., & Knittel, J. R. (2006). An objectiv e measure of service and its effect on tipping. *The Journal of Economics, 32* (2), 61–69.
DESIGN OVERVIEW : Author investigated how service quality, measured by the amount of time it took to deliver the meal, influenced the tip size. Other variables included in the analysis were gender, race (White vs. all others), and income of the served location. Some information was added to the analysis based on census information, particularly the income variable. Two delivery drivers from the same restaurant measured a ll data in this study aside from "income," which was added based on census information on the location of the delivered food. The type of payment and the magnitude of the bill were also considered in the analysis.
However, it is worth noting that this arti cle does not specify how many observations are being analyzed, or provide any information about the drivers other than state that the "personal attributes of the drivers were quite similar."
INCOME: Higher-income areas were more likely to leave better tips than lower-income areas.
GENDER: Males were found to tip marginally better than females.
CASH VERSUS CREDIT: Cash-paying customers were actually found to tip better than credit card customers, but this effect was nonsignificant when the magnitude of the b ill was considered as part of the regression equation.

Klee, E. (2004). *How people pay: Evidence from grocery store data*. Federal Reserve Board. Retrieved from http://www.newyorkfed.org/research/conference/2006/Econ_Payments/Klee_b.pdf
DESIGN OVERVIEW: Examination of household data from the Survey of Consumer Finances from 1995, 1998, and 2001. Findings indicate that the share of credit card and debit card usage has increased over the years, while the usage of checks has decreased. However, these market shares and usage rates will not apply to many tipping situations, and should only be considered for demographic groups that have credit or debit cards.
AGE: Credit card usage differed somewhat by age, such that very young heads of households and those over the age of 75 have lower credit card usage than other age groups, while debit card usage differed significantly. Debit card usage was highest among the youngest cohort and decreased as age increased.
INCOME: For both credit and debit cards, usage rates increased along with rising income brackets, indicating that more wealthy individuals are more likely to have credit and/or debit cards.

Kleven, H. J., Knudsen, M. B., Kreiner, C. T., Pedersen, S., & Saez, E. (2011). Unwilling or unable to cheat? Evidence from a tax audit experiment in Denmark. *Econometrica, 79* (3), 651–692.
METHODOLOGY: Article reports results from a field experiment conducted on Danish tax filers where tax filers were initially randomly assigned to one of two groups, where one group is subject to rigorous audits while the other is not. Subjects are then randomly assigned to three groups, where one group does not receive a notice of a future audit while the other two groups receive notices that they will be audited with different probabilities (50% or 100%). Subjects in different treatment groups are compared based on the difference in the amount of income that they report and baseline audit data, with income broken down into that income that is subject to third-party reporting (i.e., there are records kept by employers, etc., against which self-reported income can be checked) and income that is purely self-reported. The authors hypothesize that only self-reported income should be affected by past audits and the threats of future audits. Consistent with the hypothesis, the effect of the enforcement treatments on evasion is close to zero for income subject to third-party reports, but having been audited in the past and the prospect of future audits reduces evasion for self-reported income. Evasion was generally substantially higher for self-reported income. Higher marginal tax rates were found to increase evasion, though the effect was relatively small. The authors argue that the results support the importance of enforcement through third-party reporting in explaining why compliance is generally high in developing countries despite low audit probabilities and fines.

Koku, P. S. (2005). Is there a difference in tipping in restaurant versus non -restaurant service encounters, and do ethnicity and gender matter? *Journal of Services Marketing, 19* (7), 445–452.

DESIGN OVERVIEW: Thirty-five participants were randomly selected for seven different service sector businesses (245 total participants) that they indicated they had patronized within the past t hree months. Service sector business included restaurants, barbershops/hair salons, spas, golf club shops, auto detailing shops, auto mechanics' shops, and valet parking. Participants were provided a questionnaire that asked them if they tipped 15% or more of the total bill, less than 15%, or did not tip at all. They were also given a space to provide a reason for their tipping decision.

INDUSTRY/SERVICE : For analysis purposes, the researchers combined all non -restaurant services to compare against restaura nt tipping. They also combined all those who said they tipped less than 15% and those who did not tip at all. Using a chi -square test, the researchers determined that there is a difference between the reasons people tip in the restaurant industry and outsi de of it.

RACE/ETHNICITY: The researchers also compared White versus non -White respondents on tipping tendencies outside the restaurant industry, and failed to find any difference.

GENDER: They only found a marginal difference between men and women in t ipping outside the restaurant industry.

Koku, P. S. (2007). Some significant factors that influence tipping in service encounters outside the restaurant industry in the United States. *Services Marketing Quarterly, 29* (1), 23–45.

DESIGN OVERVIEW: The sampl e included 12 MBA students (6 male, 6 female) who indicated that they had used another service -sector business in addition to the restaurant industry in the past 3 months. Other service -sector businesses included spas/body massage, barbershop/hair salons, auto mechanics' shops, plumbing services, auto detailing shops, valet parking, and lawn care services. There were two sessions. All participants met in the first session for two hours and were asked about service encounters in which they tipped in the past month and what led them to do so, as well as service encounters in which they did not tip and why. The second session included 30 -minute individual sessions.

INDUSTRY/SERVICE : Using the framework of transaction cost analysis (TCA), the authors propose several factors that influence a consumer's tip in other service -sector businesses (i.e., service industries other than restaurants). From information gleaned in interviews, the authors propose that the customer's decision to tip is influenced by (1) quality of service, (2) the length of time to be served or have his or her issue resolved in an emergency situation, (3) the likelihood of repeat purchase (which is influenced by service quality), and (4) budgetary constraints.

Lynn, M. (1988). The effects of alcohol consumption on restaurant tipping. *Personality and Social Psychology Bulletin, 14* (1), 87–91.

DESIGN OVERVIEW: The author became employed as a waiter at the restaurant where the study took place. For just over a month, he recorded information for 207 dining parties, including bill size, tip amount, whether alcohol was consumed and number of drinks, customer's gender, and payment method.

BILL SIZE: A regression of tip amount on bill size indicated that tipping is strongly, positively related to bill size. The resulting equation found a *y*-intercept of .32 (32 cents) with an additional tip of 11% of bill size; this accounted for 50% of the variance in tip amount.

INDUSTRY/SERVICE: After controlling for the relationship between bill size and tip amount and a host of other variables, a hierarchical multiple regression found a significant effect for alcohol. The results indicate that alcohol (but not number of drinks) consumption increases tipping.

Lynn, M. (2004). Black-White differences in tipping of various service providers. *Journal of Applied Social Psychology, 34* (11), 2261–2271.

DESIGN OVERVIEW: A randomized telephone-based survey was conducted to determine the difference in tipping behavior among various service industries. This data was acquired by Lynn in order to conduct follow-up analysis regarding tipping differences between Whites and Blacks. Waiters, bartenders, barbers, taxi drivers, food-delivery people, hotel maids, masseuses, bellhops, and ushers at theatres or sporting events were the occupations of interest. In the final analysis, 894 respondents (811 White and 83 Black respondents) were used. Respondents were asked, "If you received good service from ____ would you tip them a percent of the total cost of the service, tip them a flat amount, or not give them a tip?" Respondents were asked this question nine times for different service industries: waiter or waitress; bartender; barber, hair stylist, or cosmetician; cab or limousine driver; food-delivery person; hotel maid; skycap or bellhop; masseuse; and usher at theater, sporting events, etc. Respondents were then further questioned about the amount they would tip if they indicated that they would tip a percentage or flat amount.

INDUSTRY/SERVICE: Waiters received the most tips among Whites, though barbers also had a high tip percentage amount among both Whites and Blacks.

RACE/ETHNICITY: Blacks are less likely to base restaurant tips on bill size than are Whites. Black percentage tippers leave a smaller average percentage of the bill than do White percentage tippers across many service contexts. Finally, Black flat tippers leave larger average dollar tips than do White flat tippers across many service contexts (e.g., bartenders, barbers, hotel maids, and masseuses).

Lynn, M. (2004). Ethnic differences in tipping: A matter of familiarity with tipping norms. *Cornell Hospitality Quarterly, 45* (1), 12–22.

DESIGN OVERVIEW : The survey results were the same as reported in Lynn's 2004 article on tipping knowledge among various racial groups in which respondents were collected by random-digit-dialing (RDD) telephone methods. Respondents were asked how much it was customary to tip waiters and waitresses in the United States with "15% to 20%" considered to be the right answer. Roughly 1,000 total completes were gained, but only 99 were from Black respondents. It is also important to note that respondents were asked about customary practices rather than their own tipping behavior.

RACE/ ETHNICITY: Overall, most Whites (over 70%) indicated that they knew the correct amount to tip a waiter or waitress, compared with only 37.4% of Black respondents. Furthermore, 12.1% of Black respondents reported that they did not know the correct amount compared with only 2.4% of White respondents.

Lynn, M. (2006). Geodemographic differences in knowledge about the restaurant tipping norm. *Journal of Applied Social Psychology, 36* (3), 740–750.

DESIGN OVERVIEW : A phone survey was conducted by Taylor Nelson Sofres using random-digit-dial sampling for a total sample of slightly over 1,000 respondents. The primary question of interest was how much people are expected to tip waiters and waitresses in comparison to how much they typically tip. The "correct" response was considered to be 15% to 20%.

RACE/ ETHNICITY: Significantly more Whites (72%) have the correct knowledge of tipping conventions compared with Hispanics and Blacks (33% of both). These effects were still significant once other variables were controlled for.

AGE: Age was initially significant, such that respondents in their 40s to 60s had higher levels of knowledge compared with older and younger respondents, but once other factors such as race, sex, education, income, metro status, and region were controlled for, it became insignificant.

GEOGRAPHY: Metro status was marginally significant before controlling for other variables but non-significant after control variables were considered. That said, the Northeast region had higher levels of tipping knowledge compared with the South region, but there were no other significant differences between other regions.

INCOME: Higher levels of income were related to higher levels of knowledge of correct tipping norms even when controlling for other variables.

EDUCATION: Higher levels of education were related to higher levels of knowledge of correct tipping norms even when controlling for other variables.

GENDER: Knowledge did not vary by sex when directly compared with men, but when other variable were controlled for, it was found that women had a significantly higher level of tipping knowledge than men.

Lynn, M. (2006). Tipping in restaurants and around the globe: An interdisciplinary review. In M. Altman (Ed.), *Handbook of Contemporary Behavioral Economics: Foundations and Developments* , (pp. 626–643). M. E. Sharpe Publishers.
DESIGN OVERVIEW: Lynn examines results from the literature for anything related to tipping. This includes the determinants of restaurant tipping, including bill size, payment method, gender, and race/ethnicity. A meta-analytic review conducted by Lynn and McCall (2000) found that 69% of the variability of dollar tip amounts within a restaurant can be explained by the bill size. Several studies' results support a "magnitude effect" where dollar tip amount increases with bill size, but percentage tip decreases. Several studies have demonstrated that patrons paying by credit card tend to leave a larger tip than those paying with cash (Feinberg, 1986; Garrity & Degelman, 1990; Lynn & Latane, 1984). Furthermore, the presence of a credit card company insignia induces higher tip amounts (McCall & Belmont, 1996). There has been some support for men leaving larger tips than women and waitresses receiving larger tips than waiters. Results indicate that patrons are more likely to provide a higher tip for a server of the opposite sex. Black restaurant patrons are more likely than their White counterparts to tip a flat amount rather than a percentage and tip a lower percentage. Studies have shown these results even when controlling for education, income, and perceptions of service quality (Lynn & Thomas-Haysbert, 2003).

Lynn, M. (2011). Race differences in tipping: Testing the role of norm familiarity. *Cornell Hospitality Quarterly*, *51*(1), 73-80.
DESIGN OVERVIEW: This study was a web-based survey from a consumer panel (Zoomerang.com) in which the aim was to test and determine if tipping knowledge mediates the relationship between race/ethnicity and tipping behavior, because no work up to this point had tested if this relationship existed. Multiple waves of invitations were sent until the desired demographics groups were gathered (100 respondents from both White and Black races, and with a separate split of those with and without a college education, 831 total observations in all).
As with previous studies, respondents were asked how much people in the United States are generally expected to tip waiters and waitresses, with 15% to 20% being considered the correct answer. Later in the survey, they were also asked about their tipping behavior for waiters and

waitresses that gave them good serv ice in order to determine not just their knowledge about tipping behavior, but also their own behavior as well.

For other industries and services, such as hotel maids and bartenders, the respondent was simply asked if that industry was generally tipped at all. Respondents who indicated that the various other services were tipped were considered as having some knowledge of the norm for that occupation.

As with waiters and waitresses, respondents were further asked how often they tipped members of the other professions.

RACE/ETHNICITY: Analyses indicate that tipping norm awareness did predict racial differences between Black and White tipping behavior for restaurant tips, both for tip type (whether a percentage of the bill was left versus a flat amount) and t he percentage left.

No racial differences were found in the tipping/stiffing of hotel maids and luggage handlers, but racial differences were found for the other investigated services.

Finally, a moderated relationship for norm awareness was also tested fo r, but this was not found to be statistically significant.

INDUSTRY/SERVICE: A few significant differences were found for other professions. Specifically, they found that norm awareness mediated racial differences in stiffing behaviors for haircutters an d pizza delivery, but not for bartenders, parking valets, or cab drivers.

Lynn, M. (2012). The contribution of norm familiarity to race differences in tipping: A replication and extension. *Journal of Hospitality & Tourism Research.* Advance online public ation. doi:10.1177/1096348012451463

DESIGN OVERVIEW : Web-based survey was sent out to members of a consumer survey panel. Response rates were not calculated as probability of panel selection is not captured. The final sample included 180 respondents after s cleaning the original data set for outlier responses (such as suggesting they gave tips over 100% of their bill) or improbable completion times...

Respondents were asked how much they would tip for one of two randomly assigned bill amounts, \$21.32 or \$46 .23, if the service was determined to be unusually good, average, or unusually bad.

Finally, respondents were also asked how much people in the United States are expected to tip a waiter for adequate to good service and given typical response options.

RACE/ETHNICITY: Controlling for age, sex, income, education, and bill size, Black and Hispanics were found to tip less, and were also less aware of the standard 15% to 20% tipping norm. Furthermore, it was found that tipping knowledge significantly affected ti p size after controlling for race, indicating a partially mediated relationship.

Lynn, M. (2013). A comparison of Asians', Hispanics', and Whites' restaurant tipping. *Journal of Applied Social Psychology, 43* (4), 834–839.
DESIGN OVERVIEW : An online survey was conducted via a large multistate restaurant, yielding 1,274 final observations after 64 subjects who refused the race/ ethnicity question were dropped from the analysis. The survey asked respondents about service and restaurant quality in addition to the size of their bill and tip size. Service quality was used as a control when observing the differences between the different racial groups.
This study asked respondent race/ ethnicity as a single -item question, as opposed to how the U.S. Census Bureau asks two questions, one for race and one for ethnicity. In this setup, respondents could indicate that they were Hispanic or Black, but not both.
BILL SIZE: Flat-dollar tips increased along with bill size while percent tips decreased in the same span.
RACE/ ETHNICITY : Hispanics tip significantly less than Whites but there are no differences between Asians and Whites. However, given the relatively low *N* of the Asian population (roughly 75 observations) the findings have to be taken with caution.

Lynn, M., & Gregor, R. (2001). Tipping and service: The case of hotel bellmen. *International Journal of Hospitality Management, 20* , 299–303.
DESIGN OVERVIEW: A hotel bellman interacted with 50 different customers while delivering one of two conditions of level of service, at a small luxury hotel. In the "limited" service condition, the bellman met customers at their cars with a cart and loaded their bags and then accompanied them to their hotel room after they checked in, opened the door, and brought the luggage to their room. They then asked guests if there was anything else they needed before collecting any tips and leaving the room. The "full" service condition included the same treatment as the "limited" condition, but the bellman also demonstrated how to use the television and thermostat, opened the blinds, and offered to get ice for the guest. The bellman recorded the guests' experimental condition, sex, apparent age, and tip following each interaction.
INDUSTRY/ SERVICE : The hotel bellman received significantly higher tips for providing the "full" service condition ($4.77) than the "limited" service condition ($2.40). The effect of increases in tips based on service condition was similar among men, women, younger guests, and older guests.

Lynn, M., & Latane, B. (1984). The psychology of restaurant tipping. *Journal of Applied Social Psychology, 14* (6), 549–561.

DESIGN OVERVIEW: In the first study, 169 groups of customers were interviewed as they exited an IHOP. Only those who paid the bill were questioned, or if two or more paid the bill, their responses were combined. Participants were questioned about party size, restaurant atmosphere, food quality, service quality, bill size, tip size, and improvements for the restaurant; respondent gender was also recorded. All servers were female.

In the second study, 4 waiters and 5 waitresses collected data for 206 dining groups over a 1-week period. They recorded the number of people on the check, number of people at the table, number of checks at the table, bill size, gender of person(s) paying the check(s), method of payment, amount left as a tip, and server's level of effort spent serving the table. They recorded this information for parties of five people or less or larger parties without a reservation because of the automatic gratuity applied to larger parties.

BILL SIZE: In the first study, the average bill size was $3.16 and the average tip per person was $.42. Customers tipped an average of 15.6% of their bill size. A hierarchical, multiple linear regression of customer's gender, party size, number of separate checks, atmosphere, service, food ratings, and per-person bill size on percent tipped was performed. After controlling for other variables, per-person bill size predicted a significant amount of variance in percent tipped. The larger the per-person bill size, the smaller the percentage tip of the total check. In the second study, the average bill size per person was $13.01 and the average tip per person was $2.01. Customers tipped an average of 15.5% of their bill size. In the hierarchical, multiple linear regression, per-person bill size was unrelated to percent tip, which the authors speculate is due to the high price of the restaurant where the study was conducted compared with that of a café in the first study where some groups only ordered coffee or a snack.

PAYMENT METHOD: In the second study, a hierarchical, multiple linear regression of customer's gender, server's gender, party size, number of separate checks, effort ratings, per-person bill size, and payment method on percent tipped was performed. After controlling for other variables, payment method predicted a significant amount of variance in percent tipped. Customers paying their checks with credit cards tipped a larger percentage of the bill than cash-paying customers (16.9% versus 14.5%).

GENDER: In the first study, using the same hierarchical, multiple linear regression of customer's gender, party size, number of separate checks, atmosphere, service, food ratings, and per-person bill size on percent tipped, after controlling for other variables, gender predicted a significant amount of variance in percent tipped. Men tipped significantly more than women (17.4% versus 9.5%). For the second study, in the hierarchical, multiple linear regression of customer's gender, server's gender, party size, number of separate checks, effort ratings, per-person bill size, and payment method on percent tipped, after controlling for other variables, customer's gender also predicted a significant amount of variance in percent tipped. Men tipped slightly more than women (15.7% versus 14.6%).

Lynn, M. & McCall, M. (2000). Gratitude and gratuity: A meta-analysis of research on the service-tipping relationship. *The Journal of Socio-Economics, 29*(2), 203–214.
DESIGN OVERVIEW: Meta-analysis conducted on a combination of published and unpublished studies that had variables concerning tipping behavior and service quality, yielding observations for 2,547 dining parties across 20 different restaurants. The unit of analysis used was the *N* of restaurants, as the authors argue that as tipping expectations and norms can vary by establishment, that is the most appropriate level for analysis. Some splits were done to determine the relationship between service quality and tipping behavior based on the metric used in the analysis and the person providing the data, as some of the relationships were based upon a server's estimation of the service quality rather than the customer's.
Ultimately, it was determined that there was a significant relationship between service quality and tips, but that it accounted for less than 2% of variance in tipping behavior. This value was stronger (almost 5%) among studies that had stronger measures of service quality. However, there was no such relationship found for measures that recorded the perceptions of servers, indicating that servers do not see a link between service quality and tipping behavior.

Lynn, M., & McCall, M. (2000). *Beyond gratitude and gratuity: A meta-analytic review of the predictors of restaurant tipping*. Working paper, School of Hotel Administration, Cornell University.
DESIGN OVERVIEW: The authors limited the meta-analysis to research concerned with the restaurant industry where the data were collected about an individual service encounter from one of three or more modes: (1) restaurant checks, charge receipts, and comment cards; (2) records kept each evening by restaurant servers; and/or (3) interviews with patrons as they departed restaurants. A total of 22 published studies and 14 unpublished studies were included in the meta-analysis. The authors meta-analyzed the relationships of tip size to bill size and of bill-adjusted tip size to 23 predictors from the tipping literature, including weather, payment method, and alcohol consumption.
BILL SIZE: The meta-analysis indicated that tip amounts were positively related to bill size. In fact, the authors found that bill size accounted for about two-thirds of the variability in tip amounts.
GEOGRAPHY: Meta-analysis results indicate that patrons left larger bill-adjusted tips when the weather was sunny.

CASH VERSUS CREDIT: Patrons left larger bill -adjusted tips when they used a credit card as their method of payment or when they received their bill on a tip tray embossed with a cr edit card company's insignia.
INDUSTRY/SERVICE: Alcohol consumption was not related to bill -adjusted tips.

Lynn, M., & Thomas -Haysbert, C. D. (2003). Ethnic differences in tipping: Evidence, explanations, and implications. *Journal of Applied Social Psychology, 33*(8), 1747–1772.
DESIGN OVERVIEW : A pair of studies were conducted to investigate racial differences in tipping. The first study was based on the data from the 1997 Speer article. The first study was based on the data from the 1997 Speer article. The second study was based on a collection of data sets based on five tipping articles that either interviewed customers after they had left their restaurant or the servers after the customers had had their meal.
The first study used the data from Speer (1997), with an *N* of about 1,000 from a telephone survey and about 100 Black respondents. The combination of data sets in the second study resulted in an *N* of about 1,800 respondents, with 94 Black respondents, 149 Asian respondents, and 113 Hispanic respondents. All the restaurants in the five studies used in the second study came from in or near Houston, Texas.
RACE/ETHNICITY: The first study showed the same results as in previous studies in that Blacks tipped less than Whites, but additional mediating analyses were conducted. Age, income, education, and tip size were al l found to be partial mediators of the race/ethnicity relationship.
The second study found that Whites left significantly higher tip sizes compared with both Blacks and Asians, but not Hispanics. Another finding of note was that Asians and Hispanics were more likely to tie the percent tip to service quality than Whites and Blacks.

Lynn, M., & Williams, J. (2012). Black -White differences in beliefs about the U.S. restaurant tipping norm: Moderated by socio -economic status? *International Journal of Hospita lity Management, 31* (3), 1033–1035.
DESIGN OVERVIEW : A pair of phone surveys were used for the analysis that used separate, but very similar, questions. One survey asked, "Thinking about tipping overall, not your own practices, how much is it customary for people in U.S. to tip waiters and waitresses?" The other survey asked, "Thinking about restaurant tipping norms, how much are people in the U.S. expected to tip waiters and waitresses?" Both questions were open -ended and results were coded into predetermi ned response options, such as "15% –20%." Tipping knowledge was considered to be either partial (in

terms of knowing that it was customary to tip waiters and waitresses) or complete (that it was customary and that 15% to 20% was the correct amount).

A measure of Socio-Economic Status (SES) was crafted based on a pair of questions, one asking income and the other asking education background. These were standardized and then averaged together to form one scale.

RACE/ETHNICITY: The significant difference between White and Black tipping knowledge was mediated by SES for partial tipping knowledge but not for complete tipping knowledge. This would seem to indicate that all low SES individuals in general are unaware that tipping is customary in certain situations, but that the "correct" tipping amount is not influenced by SES, and still seems to involve a racial component.

Lynn, M., Zinkhan, G., & Harris, J. (1993). Consumer tipping: A cross-country study. *Journal of Consumer Research, 20*, 478–488.

DESIGN OVERVIEW: The authors used information about tipping in 33 service professions across 30 different countries (Star, 1988). Each service in each country was coded as either "tipped" or "not tipped" and aggregated. The authors obtained data for four different work-related indices that they posit are related to tipping differences across countries. The study used 116,000 questionnaire responses from industrial corporation employees in 50 different countries (Hofstede, 1983). The four indices are "power distance," where a high score reflects an acceptance for hierarchical structure and a low score reflects the opposite; "uncertainty avoidance," where a high score reflects a culture that is concerned with following the rules and a low score reflects one that is willing to take risks; "individualism," where a high score is associated with a culture that is concerned with individuals' independence and a low score reflects a culture of collectivism; and "masculinity" reflects a culture whose values are primarily masculine.

GEOGRAPHY: There was a correlation of .46 between the power distance index and number of services that get tipped, indicating a strong relationship between high power distance scores and the number of services that are tipped. There was a correlation of .55 for uncertainty avoidance and tipping, indicating that tipping occurred more often in countries that were less tolerant of uncertainty. There was a correlation of -.39 between the individualism index score and tipping, indicating that tipping was more common in collectivistic countries. There was a correlation of .47 for masculinity index and tipping, indicating that tipping occurred more often in countries with masculine values. Japan was an outlier in all four analyses and was omitted.

McCall, M., & Belmont, H. J. (1996). Credit card insignia and restaurant tipping: Evidence for an associative link. *Journal of Applied Psychology, 81* (5), 609–613.

DESIGN OVERVIEW : For the first experiment, data were collected from 77 paying customers at a family restaurant; men were most frequently the paying customer (59 men and 18 women). Patrons tended to be people vacationing at a nearby ski resort. The independent variable was what type of tip tray the diner received with the check, either a blank tip tray or a tip tray with the credit card insignia of a major credit card company in the center of the tray. Servers recorded the amount of the bill, number of patrons in the dining party, the sex of the individual paying the bill, the method of payment, and the total amount tipped.

For the second experiment, data were collected from 27 paying customers from a café in a separate town from Experiment 1, whose main clientele is university students. The sample included 13 men and 12 women, and two missing cases where gender was not recorded. The methodology of Experiment 2 replicated Experiment 1 except that the credit card insignia on the tip trays was from a different credit card company.

CASH VERSUS CREDIT : In the first experiment, an analysis of covariance (ANCOVA) revealed that a credit cue significantly affected percentage tipped. Specifically, individuals that were given the tip tray with a credit card insignia tipped a significantly higher percentage (19.77%) than those who received a blank tip tray (15.48%).

In the second experiment, all paying customers used cash. Data were analyzed the same way for Experiment 2 as they were in Experiment 1. Similar to Experiment 1, the ANCOVA demonstrated a significant effect of credit cue on percentage of the bill tipped, where the presence of a credit card insignia resulted in a tip percentage of 21.91% compared with those who received a blank tip tray (17.53%). While the following two experiments did not compare tipping by method of payment used, these were the basis for a lot of method-of-payment research in the future.

McCrohan, K. F., & Pearl, R. B. (1983, August). *Tipping practices of American households: Consumer based estimates for 1979* . 1983 Program and Abstracts: Joint Statistical Meetings, Toronto, CA.

DESIGN OVERVIEW: Diary population was recruited via telephone recruitment and auto -registration listings, creating an estimate of $5.7 billion in tipping revenue. Demographic targets were based on census data. Two samples were used: 10,000 family households and an additional 1,500 nonfamily households. The sample populations were recruited via telephone recruitment and auto -registration listings. Reports were given on a quarterly basis over the course of the entire year. Families reported over a two-week span every quarter and were staggered such that there were diaries coming in from some of the sample every week. However, the nonfamily sample only reported during one quarter in the entire year.

NATIONAL AVERAGE TIPPING BEHAVIOR: Of the $72.7 billion that was spent on di ning out in 1979, 31% was considered to be spent on tipping occasions and such occasions accounted for over half of all revenue. Of this revenue, tipping behavior constituted $5.7 billion, or roughly 14.4% of tipping occasion behaviors.

After examining the data and determining what types of establishments should be classified as "tipping occasions," they determined that the true stiffing rate was somewhere around 20%, though that included some situations where people ordered hasty meals or snacks.

GEOGRAPHY: Findings indicate that tipping was higher in the northeast region of the country compared with the middle parts of the country and that metro areas tipped at higher rates.

INCOME: Very small differences were found relating to income, such that the highes t income group tipped at about 1% greater rate than the lowest income group.

CASH VERSUS CREDIT: Credit transactions tipped at a somewhat higher rate than cash transactions (1% difference), but at this point they were only used in less than 3% of all dinin g transactions.

McCrohan, K. F., & Pearl, R. B. (1991). An application of commercial panel data for public policy research: Estimates of tip earnings. *Journal of Economic and Social Measurement, 17*, 217–231.

DESIGN OVERVIEW: Authors expand on the analysis of the consumer diary data discussed in Pearl and McCrohan (1984). The diary panel of restaurant patrons now includes the years 1982, 1983, and 1984. The authors find that tipping occurs in only 29% of eating occasions, but that tipping occasions account for approximately half of all expenditures. A regression analysis was also undertaken to examine the determinants of the tipping rate (tip amount over total expenditure, for both tipping and non -tipping occasions) for a given occasion.

NATIONAL AVERAGE TIPPING RATES: Across all periods, tip rates averaged approximately 14.4% and that the average was relatively invariant across the types of eating establishments (inside, outside, or non-tipping), though stiffing behavior varied by type, with tipping type restaurants (family, atmosphere, and coffee shop) accounting for 90% of all tips.

CASH VERSUS CREDIT: Findings from the regression analysis indicate that tipping rates are higher when establishments accept credit cards.

INDUSTRY/SERVICE: Findings from the regression analysis indicate that tipping rates are higher when establishments serve alcohol.

GEOGRAPHY: Findings from the regression analysis indicate that tipping rates are higher when establishments are located in metropoli tan areas.

Morran, C. (2013, September 5). Are these the final days of automatic 18% tips at restaurants? *Consumerist*. Retrieved from http://consumerist.com/2013/09/05/are-these-the-final-days-of-automatic-18-tips-at-restaurants/.
SERVICE CHARGE: Report on the change in how IRS considers the automatic 15% to 20% gratuity in restaurants, citing the piece by Jargon (2013) in *The Wall Street Journal*. Darden Restaurants, parent company of Olive Garden, Red Lobster, and LongHorn Steakhouse, has already reported that it was going to drop the automatic gratuity policy because of this issue.

Neuman, S. (2013, September 5). IRS to count automatic gratuities as wages, not tips. *NPR*. Retrieved from http://www.npr.org/blogs/thetwo-way/2013/09/05/219290573/irs-to-count-automatic-gratuities-as-wages-not-tips.
SERVICE CHARGE: Blog post on the IRS's change in how automatic gratuities are counted. The blog post covers an original *Wall Street Journal* article on this issue (see Jargon, 2013, for original report).

Noll, E., & Arnold, S. (2004). Racial differences in tipping: Evidence from the field, *Cornell Hospitality Quarterly*, *45*, 23–29.
DESIGN OVERVIEW: Two unpublished studies, both of which were reported by servers from a large restaurant chain, were used. In the first study, approximately 100 servers were asked a variety of questions regarding supposed "tip predictors" such as race, alcohol use, and gender. The second study aimed to investigate whether servers were accurately reporting their tip sizes as that misreporting could significantly damage the results that were found in the first experiment. Two servers in the same restaurant chain (but in another state) agreed to note their tips over a two-week period. Overall, tips were recorded from 151 sets of customers.
RACE/ETHNICITY: Nearly all of the servers in the first study reported that they were aware of the differences in tipping by race. Three-quarters of the servers indicated that their Black customers were less likely to provide a tip, and when a tip was provided, more likely to tip below 15% than White customers. In the second study, the two reporting servers reported similar findings for differences between White and Black customers (though it is worth noting that outliers of tips over 26% were removed for both White and Black customers prior to analysis).
GENDER: In the first study, it was also found that male customers tipped more than female customers.
INDUSTRY/SERVICE: In the first study, it was reported that customers who consumed alcohol gave significantly higher tips than those who did not consume alcohol.

CREDIT/CASH: In the first study, servers reported that credit card customers tipped significantly more than customers who paid with cash. However, in the second study it was reported that customers who paid with cash gave marginally higher tips than those who paid with credit cards.

Papp, T. G., & Burkhammer, A. L. (2001, March). *An investigation of server posture and gender on restaurant tipping*. Paper presented at the 22nd Annual Industrial Organizati onal Psychology and Organizational Behavior Graduate Student Conference, Pennsylvania State University.
DESIGN OVERVIEW: Servers were recruited and asked to record information for 10 different dinners, alternating between squatting and standing in order t o determine how this changed tipping behavior. Servers recorded bill size, tip size, and gender of the diners. Servers were recruited from campus and by sending out survey packets to various restaurants in the area, yielding a final sample of 107 observati ons across 12 different servers. Eight of the final servers were female and four were male. Each server was instructed to record five meals squatting and five meals standing, and to only record this for small dinners or two or fewer diners.
GENDER: The only effect found was a marginally significant difference between male and female servers such that male servers had more tips, but this was the only effect that was found.

Parker, J. A., Souleles, N. S., & Carroll, C. D. (2012). The benefits of panel data i n consumer expenditure surveys. *National Bureau of Economic Research*.
METHODOLOGY: Article reviews the benefits of the panel nature of the Consume Expenditure (CE) Survey. The authors argue that repeating questions for individual respondents increases resp onse accuracy by increasing familiarity and understanding of the survey. In addition, repeat interviews allow for respondents and interviewers to check the consistency of the responses, thus further mitigating measurement error. On the other hand, requirin g repeat interviews increases the burden on respondents and thus potentially increases sample attrition and thus selection bias, though the authors argue that there is little evidence that those who drop out of the sample are different in a way that would influence expenditure. Repeat measures also help reduce noise in individual respondent expenditures that could result from irregular expenditures taking place in individual interview periods or measurement error that results from using long recall periods. When modeling expenditure, panel data allows researchers to control for individual -level unobserved fixed effects, allowing the researcher to potentially make causal inferences concerning the effect of some time - varying factors on individual expenditure. Controlling for unobserved individual fixed effects may also reduce variability in estimated effect sizes, increasing the precision of estimates. Panel data also allows the researcher to assess the dynamics of expenditure for a given household.

Paul, P., & Gardyn, R. (2001). The tricky topic of tipping. *American Demographics*, *23*(5), 10–11.
DESIGN OVERVIEW: The article used the same data source that was mentioned in the Lynn piece on differences between Blacks and Whites among various service types (2004). Roughly 900 total phone numbers were randomly called to get the survey population. The professions that were listed in the article were waiters, bartenders, barbers, taxi drivers, food delivery workers, hotel maids, skycaps or bellhops, masseuses, and ushers at theater or sporting events.
INDUSTRY/SERVICE: Waiters were tipped far more often on a percentage basis than all other listed professions (74% were tipped a percentage compared with 22% who got a flat tip), and were also tipped the highest amount when tipped by percentage (along with barbers, both at 17%). Of all other professions, the percentage of respondents who said they were tipped a percentage was much lower than that for waiters, ranging between 5% for ushers to 31% for taxi drivers and food delivery workers.
NATIONAL AVERAGE TIPPING RATE: Waiters were also stiffed the least of all the professions, with only 2% reporting stiffing behaviors. Of the other professions, only masseuses (25%), hotel maids (26%), and ushers (70%) were stiffed at rates greater than 20%, while bellhops were stiffed the least of the other professions at 10%.
GEOGRAPHY: Various regional differences were discussed, such as respondents from the Northeast region gave higher tips to waitstaff and busboys (16% to 20%, respectively) compared with other regions, but they tipped cab drivers less than other regions (21% only gave a dollar or less for cab rides compared with 13% from the rest of the country).

Pearl, R. B. (1984). *A survey approach to estimating the tipping practices of consumers*. Special report on regression analysis to the Internal Revenue Service under contract TIR-81-21, Survey Research Laboratory, University of Illinois, Champaign, IL.
DESIGN OVERVIEW: Special analysis of the 1982 data using regression. Analyses were run using both a weighted and unweighted approach in order to examine both the propensity to tip and the tipping percentage on occasions where a tip was left. Regressions using scaled weights produced somewhat better regressions and were used in the final analysis. These analyses produced R^2 values of .20 for tipping behavior, but only .13 for regressions related to the actual tipping rate. Propensity to tip was mostly predicted by whether it was for full-scale restaurants or for snack places.
GEOGRAPHY: Metro areas tipped at higher levels than nonurban areas.

Pearl, R. B., & McCrohan, K. F. (1984). Estimates of tip income in eating places, 1982. *Statistics of Income Bulletin*, *3*(4), 49–53.

METHODOLOGY: Authors attempt to improve upon prior attempts at estimating tipping income for restaurants through the analysis of a large (*N* = 10,000 households of two or more related persons + 2,800 households of one or two unrelated persons) diary panel of restaurant patrons for 1982. Respondents kept a diary where they recorded information about all eating occasions over the course of a two-week period in a given quarter. The large sample (weighted to be representative of the U.S. population in the given years) allowed for the more precise estimates, while querying customers rather than employees or managers of establishments on tipping behavior mitigated bias that may have resulted from the incentive of employees to underreport tipping income or managers to exaggerate tipping income in order to justify subminimum wages. The authors argue that the use of a diary as opposed to a survey increases the accuracy of the information provided, because details of dining occasions are recorded closer to the time of the meal. In addition, they maintain that the use of a diary lowers the probability that respondents will exaggerate the size of the tip in order to impress the interviewer.

NATIONAL AVERAGE TIPPING RATES: The results of their analysis of data from the diary imply that tips comprised approximately 7.4% of all expenditures and 14.3% of all expenditures on meals where tipping actually occurred.

INDUSTRY/SERVICE: Respondents were asked to categorize establishments in six types (family, atmosphere/specialty, coffee shop, cafeteria, fast-food and drive-in, and take-out) where the first three categories were classified by the authors as "tipping establishments." Within the tipping establishments, sit-down and specialty establishments received tips on 60% of occasions. Within this group, tips made up 12.9% of all expenditures and 14.5% of all expenditures on occasions where a tip was actually given.

Pearl, R. B., & Sudman, S. (1983). *A survey approach to estimating the tipping practices of consumers.* Final report to the Internal Revenue Service under Contract TIR-81-21, Survey Research Laboratory, University of Illinois.

DESIGN OVERVIEW: Methodology was very similar to the previous report on 1979 tipping behavior that was conducted by NPD, with a sample of 10,000 families and an additional 2,800 households containing one or two unmarried people. The study was updated to include tipping behavior not only in restaurant situations, but also in other industries, including bars, hotels, barbershops, and taxi services. In this case, each household maintained records of tipping behavior at eating places during a one-week period each quarter, with half of the sample doing this in addition to a supplementary two-week diary study over the course of two quarters that covered additional services that might get tipped (over 50 other industrie s were identified as having been tipped, but four of them accounted for 80% of such situations and were the primary focus in the report). They were also asked to provide some brief information about the type of establishment that they ate at to determine w hether it was a situation that tipping was expected in order to determine a true stiffing rate.

In order to determine if there were sources of bias in the data, an additional phone survey was conducted with 935 households during the summer months to valida te the data that was being obtained via the diary studies. The validation study reported somewhat lower tipping rates for each service, but they were within sampling error and might be due to the change in methodology between a recall-question telephone su rvey and a diary survey.

AVERAGE TIPPING BEHAVIOR: In restaurants, the tipping rate was 14.3% overall, though only one - fifth of responses came within the 14% –16% band, one-fifth of responses exceeded 20%, and another one-seventh of responses reported less than 10%. Tipping rates also decreased along with increasing household size.

The true stiffing rate was determined to be similar to levels reported in 1979, in that roughly 21.2% of tipping situations for restaurants were stiffed and about 10% of expenditu res. As noted in the prior study, it is impossible to determine which purchase included snacks and small items that might not be considered to be tip -worthy. Stiffing rates were the lowest for credit card purchases.

GEOGRAPHY: The overall tipping rate was found to be somewhat higher in the Northeast region of the country and in metro areas.

INCOME: As noted in the previous study, as income levels increased, the tipping rate also increased somewhat with greater income, but not to a large degree.

CASH VERSUS CREDIT: Credit card users gave higher tips (14.9%) than cash users (14.3%).

INDUSTRY/SERVICE: The tipping estimates that were reported for other industries, notably bars, differed substantially from independent reports. Bars and taxi services reported rece iving tips of 19% to 20% overall, while barbers received 11.6%. The average tip at hotels and motels could not be accurately assessed based on percentages, and the average tip amount was $1.89, though this amount was still higher than that reported for the other services overall.

Stiffing rates are very difficult to assess for these other noted industries because hotels might be considered to be “stiffed” even if it was simply a one -night stop at a motel, as 70% of hotel instances did not get a tip.

Pearl, R. B., & Vidmar, J. (1988). *Tipping practices of American households in restaurants and other eating places: 1985–86*. Supplementary report to the Internal Revenue Service under Contract TIR 86-279, Survey Research Laboratory, University of Illinois, Champaign, IL.

DESIGN OVERVIEW: Report on tipping behavior from 1986, including some comparisons with previous years. It was found that roughly $6.76 billion was spent on tipping in restaurants and other eating establishments compared with $6.67 billion in 1985 and $5.85 billion in 1979. However, the percentage of money spent at eating-style restaurants compared with all eating places dropped from 39% to 34% in 1986. As in previous reports, restaurants were separated into categories that were determined to be "tipping style" restaurants, though even when split in this manner the "stiffing rate" seemed higher than it should be at 30%. Given this, they were recategorized based on the main type of food in order to create a group of "high tipping–type restaurants." This category was found to have tipping incidences of more than 80% on most occasions. They also note that the estimates of tipping revenue that they produce are lower than those provided by the U.S. Census Bureau and higher than those generated by the Bureau of Economic Analysis.

In addition to their standard analyses, a regression analysis was conducted specifically using the variables and information that might be available to the IRS in order to create a framework for future use and identification of tipping discrepancies. Scaled weights and a combination of scaled and expenditure weights were used in the analysis. The run with the expenditure weights was done to try to correct for some of the downward bias that occurs when bill size increases. The expenditure-based approach accounted for a higher R^2 than the scaled approach only (16.8% versus 13.1%). Predictions using the scaled weights alone also showed somewhat higher tipping rates than were accurate.

AGE: Middle-aged and older populations had higher rates of tipping incidence compared with younger groups.

GEOGRAPHY: Regional differences were found such that the Northeast area (which consisted of the New England and Middle Atlantic Census divisions) tipped at higher rates. Nonmetropolitan areas had one of the highest negative predictive values in the analysis. Metro areas had higher rates of tipping incidence than nonurban areas and their respective census regions. Metro areas were also significant in the regression analysis.

INCOME: As in previous studies, they found some differences in tipping behavior based on income levels. In this particular report, they found that tipping incidence was higher with higher socioeconomic statuses. The difference in tipping rate between the highest and lowest income group was only about 1%, so the range in this type of tipping behavior was not too great.

EDUCATION: Education had a similar effect on tipping incidence as did income, but had a great range of tipping rates. Tipping rates were also 1.5% higher among the hi ghest education group compared with the lower groups.
CASH VERSUS CREDIT: Credit cards had the largest coefficient in the regression analysis, showing that credit card users had higher tip percentages than those who paid with cash.
INDUSTRY/SERVICE: Establishments that served alcohol were not found to be as important to the regression analysis as had been found in previous reports.

Rind, B. (1996). Effects of beliefs about weather conditions on tipping. *Journal of Applied Social Psychology*, *26*(2), 137–147.
DESIGN OVERVIEW: In the first study, 266 adult hotel guests (181 males and 85 females) were put into four conditions. A room -service server reported one of the four weather conditions (sunny, partly sunny, cloudy, or rainy) when asked or vol unteered the information (if the guests didn't ask) while delivering food or drinks. He always reported temperatures in the 50s. The windows of the hotel rooms were soundproof and dark -tinted that gave the impression it was cloudy even under sunny conditio ns.
In the second study, 205 adult hotel guests (115 males and 90 females) were randomly assigned to four conditions. A room -service server reported one of the four weather conditions (cold and rainy, cold and sunny, warm and rainy, warm and sunny) when as ked or volunteered the information (if the guests didn't ask) while delivering food or drinks.
GEOGRAPHY: For the first study, a linear contrast analysis revealed a significant positive association between believed weather conditions and tipping. Tipping p ercentages improved as the conditions went from rainy (18%) to cloudy (24%) to partly sunny (26%) to sunny (29%).
For the second study, an ANOVA demonstrated that hotel guests in the sunny condition tipped significantly higher percentages than those who we re told it was rainy. However, there was no effect for the temperature conditions.

Sanchez, A. (2002). The effect of alcohol consumption and patronage frequency on restaurant tipping. *Journal of Foodservice Business Research, 5* (3), 19–36.
DESIGN OVERVIEW: A waitress at a steakhouse restaurant collected data for 164 tables during dinnertime over a three -month period; however, only 138 tables (158 parties) were included in the analysis. The waitress recorded several variables of interest, including group et hnicity, group size, number of parties (number of checks), party size, customers' and paying patron's ages (ages estimated), customers' and paying patron's gender, number of alcoholic beverages (for the party and

for paying patron), food bill per party, bill size per party, tip amount per party, and payment method. Several other variables, including the number of children per party and patronage frequency, were recorded and analyzed.

RACE/ETHNICITY: For analysis purposes, customers were either identified as Caucasian or non-Caucasian. Ethnicity did not have any significant effect on tipping behavior. Caucasians tipped slightly less ($7.42) than non-Caucasians ($7.49).

AGE: Results indicated that estimated age of the paying patron by the server was a good predictor of tips. Older, paying patrons tipped more than those paying patrons judged to be younger.

CASH VERSUS CREDIT: Paying patrons' choice of payment method (i.e., cash, check, or credit) did not have any relationship with the total tip amount. Those patrons who paid with cash or check tipped slightly more ($7.49) than those using credit ($7.42).

INDUSTRY/SERVICE: Consumption of alcoholic beverages was found to significantly affect the tip amount. Tips from paying patrons who had one alcoholic drink ($10.19) or more than one ($9.52) tipped significantly more than those who did not drink an alcoholic beverage ($6.44).

Schwer, R. K., & Daneshvary, R. (2000). Tipping participation and expenditures in beauty salons. *Applied Economics*, *32*, 2023–2031.

DESIGN OVERVIEW: A stratified, convenience sample of 317 respondents was selected for this survey. This sample included a mix of respondents from banks, university staff and students, government employees, and customers of barbershops and beauty salons. Furthermore, the survey was conducted over time periods during the spring and summer of 1995. Questions on the survey dealt with patronage, what barbershops or beauty salons they go to, important qualities in the salon or barbershop they go to, and various demographic and socioeconomic questions.

Analyses were conducted using a combination of probit and Tobit regressions. Two Tobit regressions were used, a censored version as well as a truncated run. The truncated, two-step Tobit model showed the better fit of the Tobit models.

SERVICE/INDUSTRY: Overall, while Post (1992) recommended tipping 15% to 20% for hair salon/barbershops, it was determined that all customers tipped at 8% of their bill, and 9% when customers who left no tip were excluded.

INCOME: Income was included in the analyses, but significant findings were only discovered in the probit analysis. In both cases the dummy variables showed marginally significant findings.

AGE: Results from the truncated Tobit analysis indicate a marginally significant finding for tipping behavior, such that tipping rates decrease with the age of the respondent, though no such significant finding was discovered in any of the other data runs.

GENDER: Mixed findings regarding gender were found between the probit and truncated Tobit models. The probit model showed a marginally higher tipping total from women than men, but an opposite finding was reported in the truncated Tobit analysis.
RACE: White respondents were found to tip marginally more in only one of the three models (the censored Tobit) and race was generally found to be a nonsignificant variable.

Seiter, J. S., & Weger, H., Jr. (2013). Does a customer by any other name tip the same? The effect of forms of address and customers' age on gratuities given to food servers in the United States. *Journal of Applied Social Psychology*, *43*, 1592–1598.
METHODOLOGY: A field experiment of diners (N = 142) at two Utah restaurants was conducted to examine the effects of differences in how servers addressed customers (first name, Mr./ Mrs., etc.) on tip rate. A regression analysis was conducted that included form of address effects, customer age, and the interaction between age and form of address. Data was collected by three student/servers.
AGE: In the regression model without the interaction (i.e., just form of address and age), customer age had a negative association with tip amount, but the estimated relationship was not found to be statistically significant at the 5% level, but was at the 10% level (p = .09). This negative relationship was stronger when the customers were addressed by their first name.

Simpson, H. (1997). Tips and excluded workers: The New Orleans test. *Compensation and Working*, Bureau of Labor Statistics, 32–36.
DESIGN OVERVIEW: Data was gained from "BLS field economists" in face-to-face interviews for the most part. Of the 359 establishments that were sampled, 77% provided some data, but only 11 provided tipping data, indicating that the findings in this article are to be considered as preliminary without any significance testing. Besides information regarding the number of tipped workers at the establishment and the dollar amount of tips collected, the BLS workers also gave a rating for their confidence in the data that was provided. However, while the majority (82%) of the data for "hours worked" was determined to be good, only 55% of the tip data was considered to be good, and 27% was considered "poor." This indicates that the data in the article might be flawed and underscores the difficulty of obtaining reliable tipping data.
SERVICE/INDUSTRY: Of the occupations that met publication criteria (certain number of workers from a certain number of establishments at least), waiters had the highest amount of average tips per hour ($6.10), followed by hostesses ($5.73), bussers ($4.86), and bartenders ($3.70).

GEOGRAPHY : The article reports that tipped employees would underreport their tips during the busy months and overreport during the slower months in order to balance things out for their bosses and create less hassl e. Similarly, the months when data was collected (July –August) were considered to be slower tourist months, so the data might be skewed somewhat by that.

Speer, T. (1997). The give and take of tipping. *American Demographics, 19* (2), 51–54.

DESIGN OVERVIEW : Random telephone survey of roughly 1,000 adults in 1996. Respondents were asked what the largest determining factor was regarding their tipping behavior, and service was often claimed as the most important thing, though this percentage was smaller among n on-restaurant services.

INDUSTRY/ SERVICE : Roughly 28% of respondents indicated that they never tipped the individual in the hotel who replaces their towels and bed sheets. Also of note was that 36% of respondents indicated that they always carried their ba ggage at hotels and airports, and were thus unable to answer any questions about tipping this particular profession. Similarly, roughly half of adults reported that they don't use taxi cabs or limo drivers, so they were unable to answer any such questions about tipping behavior. Finally, 40% of respondents indicated that they are never served by bartenders.

Also worth noting is that this article has a chart that indicates the percentage of respondents who indicate specific tipping percentages for a number o f different industries.

INCOME: Higher-income ($50,000 or higher) individuals reported that the reason they tipped was that they tipped to help some individuals (notably parking valets, luggage handlers, and taxi drivers). Lower-income individuals were les s likely to tip at all because they reported that the bill should reflect the full cost of the service, though this behavior does not extend to waiters.

GEOGAPHRY : Southerners were more likely to say they would never tip for some services, mostly taxi drivers, waitstaff, and barbers, while Midwesterners were the most likely to say that they would never tip parking valets, bartenders, maids, and luggage handlers. Northerners tipped the highest of the groups when split by region, or reported as much.

GENDER: In this study, women were reported as more likely to leave a tip than men, particularly when it comes to services other than taxis or waitstaff. Women are more likely to report that they tip based on the impact that it has on others when compared with men.

Star, N. (1988). *The international guide to tipping: When, where, and how much to tip in the U.S. and around the world* . New York, NY: Berkley Books.

DESIGN OVERVIEW : Star's book discusses cross -country differences in tipping. Specifically, the author describes expectations and norms for tipping across 38 professions in 34 different countries. The 38 professions cover a diverse set of service -related professions including restaurant jobs (e.g., severs, bartenders, hostesses, etc.), guides, hotel staff, and hair stylists. According to Lynn, Zinkhan, and Harris (1993) that had correspondence with Star, her tipping suggestions and summaries were primarily based on questionnaires sent to hotels, national railroads, resorts, restaurants, tour groups, and so o n in each of the 34 countries.

Thomas-Haysbert, C. D. (2002). The effects of race, education, and income on tipping behavior. *Journal of Foodservice Business Research,* 5(2), 47–60.

DESIGN OVERVIEW : Phone surveys were conducted on a population of 1,005 respondents. The phone survey was conducted by Market Facts for *American Demographics* and methodology of the phone survey is discussed in greater detail in another article (Speer, 1997). Questions were asked regarding whether respondents tipped various service -industry workers such as servers, bartenders, taxi drivers, parking attendants, and luggage handlers, and why they tip or did not tip.

INDUSTRY/SERVICE: Luggage handlers were tipped the most (98% said they always tipped this group, followed by servers, parking attendants, taxi drivers, and bartenders.

INCOME: Income was found to significantly affect tipping behavior and when used as a dichotomous controlling variable it nullified the influence of r ace on tipping behavior.

EDUCATION: Same effect as income was found in that it is significantly related to tipping behavior and when used as a control it nullifies the effect of race on tipping behavior.

RACE/ETHNICITY: White respondents tipped every categ ory of worker significantly more often than Black respondents, but this effect was nonsignificant once education and income levels were considered for all service workers except for taxi drivers. However, Black respondents were more likely to indicate that service quality was more important to them than White respondents and that they tipped more to ensure better service in the future. Blacks were also more likely than Whites to indicate that they did not tip because they felt that it should be included in the bill. Black respondents reported that they tipped less than Whites but this effect was nullified when income and education were incorporated into the model.

Appendix C – Search Engines and Search Terms

Table 5. Search Engines

Search Engine	Description
University Library System	Online database of journal articles maintained by local DC Metro University.
Google Scholar	Google search engine that produces links to both gated and ungated scholarly articles.
JSTOR	Archive of peer-reviewed articles published in academic journals.
Social Science Research Network (SSRN)	Archive of social science working papers.
Business Source Complete (EBSCOhost)	Database containing archived peer-reviewed articles published in business-related journals.
ABI/ INFORM Complete (ProQuest)	Database containing peer-reviewed articles published in business-related economics, business, accounting, and marketing journals.
Accounting & Tax (ProQuest)	Database containing peer-reviewed articles published in high-impact accounting, auditing, tax management, and tax law journals, as well as trade publications.
PsycINFO	Database of peer-reviewed behavioral science and mental health articles.

Table 6. Search Terms

Search Term	Themes
Gratuity, tipping, tip giving, stiffing behavior, tip reporting	GENERAL TIPPING, NATIONAL AVERAGE TIPPING RATES
Internet, mail, and mixed -mode surveys: the tailored design method.	METHODOLOGY
Regional, urban versus rural, metropolitan tipping differences, holiday differences, seasonal effects, tourist tipping	GEOGRAPHY
Income, education, age, gender, SES, salary tipping differences/ restaurant tipping differences	INCOME, EDUCATION, AGE, GENDER
Black-White/ Asian/ Hispanic/ racial tipping differences	RACE/ ETHNICITY
Tipping knowledge, tipping norms	TIPPING KNOWLEDGE
Service charge law change, mandatory service charge, mandatory restaurant tips	SERVICE CHARGE
Tipping differences by industry, non-restaurant tipping, tipping in services industries, alcohol and tipping	INDUSTRY/ SERVICE
Method of payment tipping, credit card/ cash tipping, cash differential	CASH VERSUS CREDIT

FORS|MARSH
GROUP

IRS Tipping Report on Cognitive and Usability Testing

Prepared for Internal Revenue Service

Prepared by Fors Marsh Group LLC

January 2015

Version 1.1

Table of Contents

Executive Summary of Findings and Survey Changes

The primary aim of the IRS is the lawful collection of taxable revenue in the United States. This mission is complicated by the existence of certain business and personal services that engage in substantial cash-based transactions or have other "off-the-books" income. Service industries where tipping occurs is the kind of economic activity that poses a challenge for tax administrators. Current IRS estimates of tipping income are based upon primary research conducted decades ago. In this case, the estimated tipping rate in the restaurant industry that was reported in 1982, 14.5%,[1] has likely risen over time because of increased use of electronic payment methods such as credit and debit cards, the use of which have been shown to result in higher tip rates than cash payments. Furthermore, much of the previous research on this topic was limited in scope, focusing only on the restaurant industry. Though some estimates exist for other, commonly tipped services in the United States, such as taxis or barbers[2], the literature existing on such services is relatively scant and needs expansion. One of the primary aims of this project is to determine the frequency of use of other tipped services by respondents.

To remedy these issues of scope and accuracy, the IRS began a series of task orders aimed at determining what scholarly literature had uncovered concerning tipping-related behaviors in the United States prior to launching a large consumer expenditure survey effort to revise these estimates. A report from Fors Marsh Group[3] (FMG) identified several themes and key findings in the research that had been conducted on tipping-related behaviors. In addition, this report discussed various methodology approaches that could be used for a large-scale survey effort to update these estimates, including advantages and drawbacks to each approach.

To ensure that the findings of this survey effort produce accurate tipping estimates, it is imperative that the final survey instrument's language and structure be as refined as possible in order to maximize accuracy and scope and minimize confusion and burden. Two drafts of a proposed survey instrument were constructed by IRS and FMG researchers for usability and cognitive testing. These proposed variants included all services that were originally considered as commonly tipped but varied in how respondents were asked to record their recent expenditures. All elements of the instrument had to be rigorously examined, including the final list of services to include, how the survey should be organized and worded for minimum confusion, cross-platform compatibility, and recall length. To accomplish these goals, testing was conducted in two phases. The first phase consisted of a week of cognitive testing that focused on the users' comprehension of survey focus, wording, and organization. The second phase consisted of two weeks of usability testing that focused more on device usage (e.g., desktop, smartphone, tablet) and recall length (i.e., 1-, 3-, and 5-day periods). In both phases, changes to the survey were made in a rapid, iterative fashion at the end of each week and sometimes each day. Each week of testing had 12 unique participants, for a total of 36 participants in total. Participants that were involved in the first week of cognitive testing were not eligible to join the usability testing phase.

---

[1]Pearl, R. B., & McChrohan, K. F. (1984). Estimates of tip income in eating places, 1982. *Statistics of Income Bulletin, 3*(4), 49–53.
[2] Pearl, R. B., & Sudman, S. (1983). *A survey approach to estimating the tipping practices of consumers*. Final report to the Internal Revenue Service under Contract TIR-81-21, Survey Research Laboratory, University of Illinois.
[3]*Estimating Consumer Tipping Behavior: Review and Recommendations*, February 2014. Internal report prepared for the Internal Revenue Service by Fors Marsh Group under contract TIRNO-13-Z-00021-001.

Although numerous edits and changes were made to the original draft (See Appendix A for original draft), the findings did not indicate that any questions had to be fundamentally altered or that there were severe comprehension issues. Most language edits were made during cognitive testing, while usability testing focused on device preference and recall findings. The primary findings and improvements are listed below.

- Survey Language and Tipped Services
    - Added a new category for “moving and household maintenance” services, and revised services for all major categories based on participant feedback.
    - Revised language and examples provided for non-monetary gift questions because of respondent confusion. Some could not easily grasp the concept or understand when such a situation might occur: “I can’t think of a scenario that a non-monetary tip would occur.”
    - Improved service-specific instructions were included to help respondents understand how to fill out the survey accurately.
- Survey Construction and Device Usage
    - Determined which of the proposed survey variants was preferred by respondents. Consensus was toward the version that asked for records service by service, with many stating reasons such as the following: “I like this one (Version 2). Having everything listed out there (in Version 1) is just a lot there and I feel that people might skip over some of the lesser expenses.”
    - Crafted draft survey instrument on two survey platforms to find one that was more compatible for completion on mobile devices.
    - Conducted usability testing on three web-capable devices (smartphone, tablet, laptop) and found no significant issues with survey completion on any device.
    - Most respondents reported they would complete the survey on a computer or smartphone.
- Recall Accuracy
    - Discovered that all respondents often rely on estimation heuristics, even for short periods of recall, using language such as “I usually tip 20% and the bill was about $20; therefore, I probably gave a $4 tip.”
    - Varied recall time during usability testing to provide recommendation for survey completion. Determined that 1-day recall is preferable to maximize accuracy for all estimates.

Cognitive Testing

Design and Lab Setup

Goals: The goals of the cognitive test of the survey were to (1) determine the optimal paradigm for the survey (i.e., grouping categories as one item or breaking them out into separate items); (2) identify and repair problematic survey language and instructions; (3) ensure the response options accurately captured respondents' tipping behaviors and service-related expenditures. These goals aim to optimize the user experience and statistical reliability and validity of the survey.

Recruitment: We contacted FMG panel members about the participation opportunity. Panel members were recruited in and around the Washington, DC, metropolitan area through traditional and social media marketing. People interested in participating in the study completed a web screener. The recruiter then rescreened respondents over the phone to confirm the information they submitted. All participants were 18 years of age or older. Participants were scheduled for the interview at a time convenient for them. A total of 12 participants were interviewed during cognitive testing. The interviews were all 45 minutes in length or less.

Method: The moderator read an introduction to participants which informed them about the planned activities during the test as well as their rights as participants. Participants were either first provided with a combined version (Version 1) or expanded version (Version 2) of the survey (see Appendix A for original survey materials with both proposed survey versions). The combined version asked participants to recall tipping expenditures for each of the broad categories before asking them to select an appropriate subcategory for the service received. The expanded version broke out the survey into individual categories. After completing one version of the survey, participants reviewed the other version and gave their feedback. The order in which the surveys were presented was rotated across participants.

The moderator facilitated the cognitive test with the participant in the same room. As the participant responded to the survey, the moderator asked probing questions about the survey language and response categories. After the survey was completed, participants completed a satisfaction questionnaire that asked them questions about their experience completing the survey. The moderator then conducted a debriefing interview that went into more depth about participants' experience and perceptions about the survey.

Cognitive Test Findings

Optimal Paradigm: A primary goal of cognitive testing was to determine the optimal paradigm for the survey. Version 1 consisted of grouping all of the service categories into a single item, asking respondents to provide details for any expenditures or payments made at a restaurant or other prepared food/ drink service, casino, personal grooming or beauty service, moving or household cleaning/ maintenance, hotel/ motel, or a taxi/ limousine/ shuttle service prior to selecting the appropriate service from each broad category. Version 2 consisted of breaking out the service categories into separate items. For example, Version 2 asked respondents to provide details for any expenditures or payments made at a restaurant or other prepared food/ drink service. If they responded that they had, they received a follow-up question asking them to select the appropriate service type and fill out their transactions. This question was repeated for each of the service categories.

Version 2 of the survey was clearly preferred. 10 of 12 participants preferred Version 2 (separated items), one participant preferred Version 1 (combined item), and one participant was indifferent.

- "This [Version 2] flows a little bit easier. It's easier for my brain to think about it."
- "I like this one [Version 2]. Having everything listed out there [in Version 1] is just a lot there and I feel that people might skip over some of the lesser expenses. With this version [Version 2], it goes over every category so I think it would be more difficult for someone to skip over any expenses. This one [Version 1] is just overwhelming. It's too much information at once."
- "This first one is totally confusing. I think that the biggest problem might be that it could make people feel stupid. This one [Version 2] is already laid out, it's already delineated. It just helps the individual focus and answer what is being requested."

Language and Response Option Findings

Restaurant or Other Prepared Food Service: Numerous minor edits were necessary for the original language and service options in this category. Participants considered "Bar" to be a separate response option from "Full-Service Dining (e.g., traditional restaurants)." One participant said, "If it said something like 'bar and grill' or something like that, then I think it would be OK." As a result, the response option "Bar" was added to the category, which itself was changed from "Restaurant or other prepared food service" to "Restaurant or other prepared food/drink service."

Additional testing determined that more service categories were needed. One participant expected to see "Coffee Shops" as a separate response option in the "Restaurants or other prepared food/drink service" category. This participant said, "I would put it in its own category because it's so ubiquitous." As a result, "Coffee Shops" was added as a response option. Two participants expected to see "Food Truck" as a response option under the "Restaurant or other prepared food/drink service" category. As a result, "Food Cart/Truck" was added as a response option. Finally, the category "Self-Service/Cafeteria/Buffet" was added because some participants indicated that such food options would not fit well into the previous categories.

Other participants commented it was odd to see "Ice Cream" as a response option under the "Restaurant or other prepared food/drink service" category. One participant said, "Ice cream—that sounds weird." Another said, "I find it interesting that they put the 'Ice Cream' one in there. I mean, I guess there are ice cream stores out there, but I would have never thought to have put that in there." As a result, "Ice Cream" was collapsed with "Smoothie Shops" in order to reduce the number of service categories.

Hotel/Motel: Participants had difficulty reporting tipping behaviors for the "Hotel/Motel" category. If they tipped for housekeeping or another service in a hotel/motel, they were unclear what to indicate in the "Amount you paid for total bill payment" section. Participants said they thought they should record the amount paid for the room or the total cost for the stay at the hotel. One participant said: "I don't see anything that would be easily identifiable as a payment for the room. After I paid $130 for the room, I would leave a little money for the housekeeping. I didn't pay for the housekeeping, per se, but then I tip on the housekeeping." This issue was reexamined and a resolution was implemented during usability testing.

Another issue that arose was that participants expected to see dining options under the "Hotel/Motel" category for situations when the hotel had a restaurant, bar, or similar option. It was decided that services that could be encountered in multiple service categories should be repeated under each category. As a result, service options for "Full-Service Dining (e.g., traditional restaurant)"; "Bar"; and "Self-Service/Cafeteria/Buffet" were added to the "Hotel/Motel" category. Our concern for capturing duplicate responses was outweighed by the possibility of not capturing this service-related expenditure. To reconcile this concern, respondents were instructed not to record service-related expenditures that were previously recorded in other survey items. Duplicate responses recorded can simply be determined during analysis.

Hair Stylist/Barber becomes Personal Grooming, Beauty, or Massage Services: Numerous additional services were added to the original two that were proposed for this service category after participants identified many other beauty-related services that could receive tips. Additional services were added for "Manicurist/Pedicurist," "Massage Therapist," "Waxing/Hair Removal," and "Facial/Skin Care." Furthermore, to better reflect the new services, the category was renamed "Personal Grooming, Beauty, or Massage Services."

Moving or Household Maintenance: Participants expected to see a category that captured tipping activities for people in moving, cleaning, plumbing, and repair occupations. As a result, the category "Moving or household maintenance" was added to the survey. The services added to this category during cognitive testing were "Professional Movers," "Maid or Cleaning Service," "Lawn/Gardening Service," and "Handyman/Repairman."

Casino: After "Bar" was added to the restaurant category, one participant also expected to find the response option "Bar" under "Casino." This participant explained how a person could have an expenditure at a bar or restaurant while visiting a casino. When participants were asked what response options they expected to see under the "Casino" category, others said that they expected to see "Bar" and "Restaurant" there. Ultimately, the three food services that were added to the "Hotel/Motel" category were also added to the "Casino" category.

Taxi, Limousine, Rideshare, or Shuttle Service: The only edit required for this category was changing "App-Based Taxi" to "Ride-Share service (e.g., Uber or Lyft)." Originally, respondents were not clear that "App-Based Taxi" meant to refer to businesses such as Uber or similar services. Participants indicated that they would record the payment type for Uber as a "credit" or "debit" transaction. Participants were unaware that Uber charges an automatic gratuity, so this was not likely to be captured in the survey. Participants consider a "Smartphone credit or app" payment type to be money that has already been loaded into the smartphone or app. Participants explained this payment type to be "Google Wallet." When the moderator asked participants to describe a smartphone or app-based payment, one participant said, "I know you have apps where you put in your information and when you use it, it just takes the money from your account." Although these responses indicated an understanding of the concept, such records might need to be examined during the pilot study to ensure that this option is being selected in conjunction with appropriate services.

Other Findings

Likelihood to Use Receipts and Financial Statements: When the moderator asked participants, "Would you look up any records/receipts or complete it on the spot from memory?" participants provided varying responses. Five of 12 participants indicated that they would check their receipts and bank statements if they had difficulty recalling the transactions from memory. Two of 12 participants said whether they checked their receipts and statements would depend on the incentive. One of these participants said: "Fifty-fifty, maybe. It would depend on convenience of doing so and the incentive." The other participant similarly said, "If there was an incentive, I would go through my accounts to try and fill it out." Three of 12 participants indicated that it would depend on the purpose of the survey. One of these participants said: "If I thought it was for a specific purpose, I would check my bank statement. If it was clear that they were investigating the tipped minimum [wage] versus non-tipped minimum wage, then I would be more diligent in how I reported information." Two of 12 participants said they would only recall transactions by memory regardless of the incentive or purpose of the survey.

Accuracy of Recalling Payments: Based on the records given by participants during cognitive testing, survey respondents are likely to use heuristics for calculating payment and tip amounts. For example, one participant indicated that he used a 20% estimate when calculating the tip amount. Another participant explained that he knows he tips around or about 20%, so he just moves the decimal over and doubles it. Such responses indicate that respondents are not likely to provide precise tipping amounts and will try to calculate the amount they believe they tipped using their recollection of the bill, rather than an actual recollection of the tip amount itself.

However, participants indicated that transactions that are frequently incurred for the same amount are more likely to be accurately recalled. For example, one participant explained that his barber always charges him $16 and he always tips her $4 to bring the bill to an even $20. Similarly, another participant explained how he always pays the same amount for the barber, so this transaction was easy to recall. This same participant explained that he just remembers what the total bill was for restaurants and always tips the same percentage of the bill.

During these sessions, the moderator asked all participants, "Regarding the tipping expenditures that you have just recorded, how many days prior would you think you could accurately remember (within a half dollar) your tipping expenditures." Participants commented that it was easy to recall if the expense was incurred in the last day. Most commented that they could remember accurately up to 3 to 4 days ago, while a few indicated that they could remember expenses incurred from a week ago or longer. The longer the recall period, the more likely that participants would use heuristics to estimate tipping amounts. For example, one participant said: "You know something? It wouldn't really be that difficult for me to recall since I usually tip about 15%." Such responses might indicate that a short recall time is required for use in the survey in order to minimize respondents' use of such estimation tactics as much as possible.

Non-Monetary Tip: The term "Non-monetary tip" and the initial description for it were confusing to most participants. All participants in the cognitive testing round indicated that they had never given a non-monetary tip and most did not know what the term meant before they read the description. One participant said, "I can't think of a scenario that a non-monetary tip would occur." One participant was offended that the IRS was asking about a non-monetary payment. She said, "That feels very IRS, very in your face.... 'Pay me because I say to.' That feels invasive." The participant went on to explain that she does provide a gift to her hair stylist; however, this is not done as part of any transaction and is a part of a personal relationship, not just a service one. Because of this, the description of "Non-monetary tip" was streamlined to remove excess verbiage and to emphasize that items given as personal gifts were not meant to be recorded.

Splitting and Separating Payments: When asked, "Have you ever left a tip for someone and split the tip across payment methods, such as cash and credit card?" all 12 participants indicated that they could not recall a time when they had split a tip across payment methods.

Participants were also asked how they would record multiple expenditures at the same establishment. The moderator asked participants, "If you filled out this survey for an occasion where you went to a restaurant and had a drink at the bar before going to your table for your meal, how would you record that in this survey?" Ten of 12 participants indicated that they would record these as separate transactions. The remaining two indicated that they would record this as one transaction because it occurred at the same establishment. Instructions were added to the survey to explain that such situations should be recorded as separate transactions.

Influences on Respondent Behavior: Additional debriefing questions were asked of participants to determine if they could guess the intended use of the survey and whether such knowledge might influence their likelihood to report their transactions accurately. The moderator asked participants, "What do you think the purpose of this survey effort is?" Six of 12 participants correctly assumed that the study was being conducted for the IRS to determine tipping rates for different industries. After those six participants responded correctly, they were asked a follow-up question: "Would that knowledge make you more or less likely to fill out the survey accurately?" All six participants indicated that knowing the purpose of the study would either increase or have no impact on the likelihood that they would complete the survey accurately. Furthermore, these participants wanted more detailed information about the purpose of the study. One participant said: "I would be more inclined to try harder if I knew that it was part of some kind of decision or policy. If I saw this was from...I don't know, Verizon, I would not be very inclined to try very hard to remember." Another participant who wanted more information said: "I guess it depends on what they are going to use that information for. I would still fill it out."

Finally, the moderator asked, "Have you ever worked in a job that receives tips for your service?" If the participant responded "yes," the moderator then asked, "Does that influence how much or how you tip?" Four of 12 participants indicated that they had worked in a tipping-based occupation. Only one of these four participants said that he thought that working in a tipping-based occupation resulted in him tipping more generously.

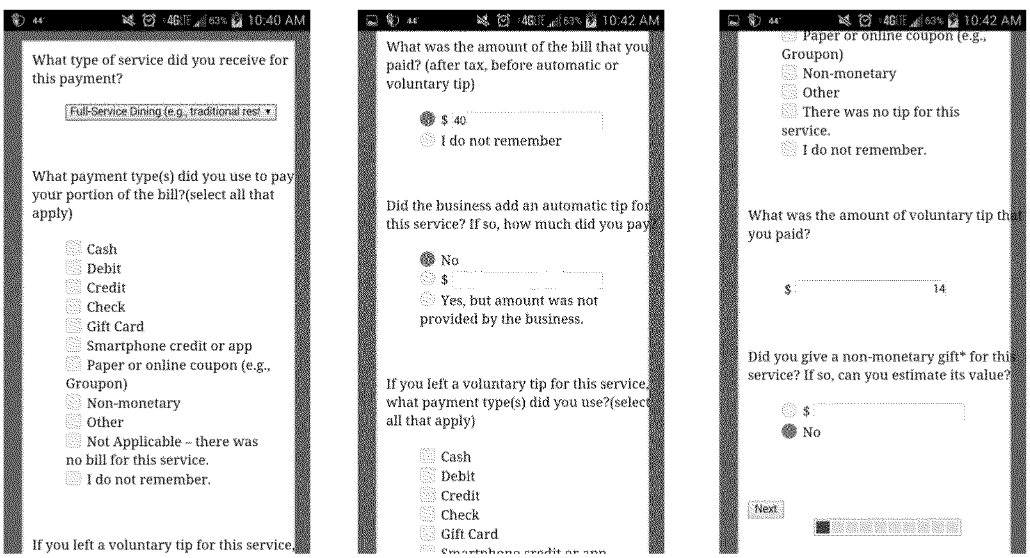
Usability Testing

Design and Lab Setup

Goals: In addition to any wording edits that were found to be necessary based on respondent confusion, there were two new goals in the usability testing phase: (1) test the survey on multiple devices to ensure that respondents are able to complete the survey on common web-based devices, and (2) examine responses across different recall periods in order to make a recommendation about the recall frame used for the pilot study.

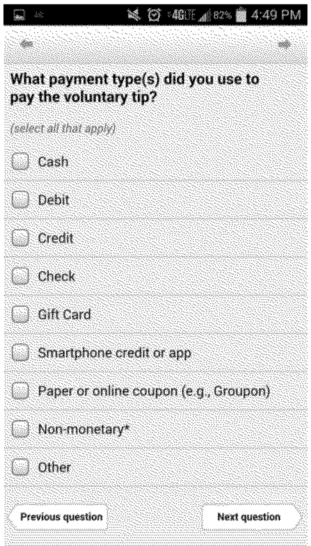
Survey Design and Platform Choice: Prior to usability testing, two different survey platforms were used to create a draft of the survey in order to determine which one might best allow for completion on mobile devices. Survey drafts were created on both Verint and Confirmit systems, two commercial survey products frequently used by FMG. The Verint platform had participants enter numerous pieces of information for a transaction on one page, while Confirmit had participants enter one piece of information on each page.

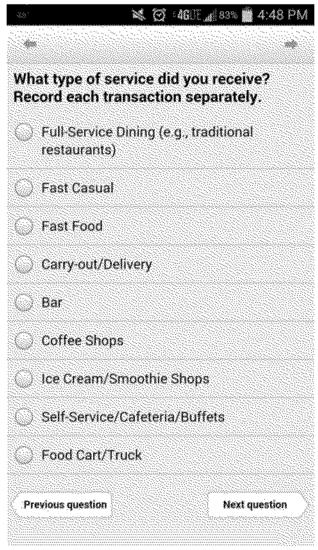
Mobile screen pictures of the Verint survey platform.



In order to facilitate completion of the survey on mobile devices, particularly smartphones, the decision was made to set up the survey on the Confirmit system. Because there are numerous pieces of information that have to be entered for each transaction, the Confirmit survey could place additional burden on the participant. However, this was judged to be an acceptable trade-off given the formatting issues present on mobile devices with the Verint survey.

Mobile screen pictures of the Confirmit survey platform.





The moderator was present to assist with any technical difficulties but tried not to assist participants in completing the survey, and often asked participants to complete the survey as if the moderator was not in the room. A few minor technical difficulties required moderator intervention, but these did not represent difficulties with the survey language or setup. Specific difficulties that arose are discussed under "Device-specific difficulties."

For each session, participants completed the survey three times, with differing lengths of recall time and different devices. While the device rotation was randomized, recall time increased gradually from 1 day to 3 days and then 5 days. To minimize burden, participants were asked to record any additional expenditures they had made during the expanded time frame and not rerecord expenditures they had listed in earlier recall periods.

Recruitment and Lab Setup: 12 participants were recruited each week for both weeks of usability testing, as was done in the cognitive testing phase. Two rounds of usability testing were conducted that took place during weeks 2 and 3 of the overall testing period. In this section, round 1 of usability testing will refer to week 2 of overall testing, while round 2 of usability testing will refer to week 3 of overall testing. Recruitment incentives and participant burden were unchanged from the cognitive testing phase.

After participants were briefed of their rights, the session began. Each participant completed the survey three times, with differing lengths of recall time and on different devices. Participants completed surveys with 1-day recall before completing surveys with 3- and 5-day recall. For the sake of time, participants were instructed only to record new transactions during the longer recall times. Unlike recall time, device order was randomized among participants in order to get novel reactions to the survey on each device. Three devices were used: a Windows laptop computer, an Android smartphone, and an iPad tablet.

Usability Testing Findings

Device Preference: The majority of participants across both rounds of testing indicated that they preferred completing the survey on the personal computer the most out of the three devices. However, eight participants during week 1 said that they would likely complete the survey on their smartphone. The stated reason was often that they were not at home to use their personal computer or that they always had their phone with them. Furthermore, although participants said that it was more difficult to complete the survey on the smartphone, they said that completing the survey on all three devices was easy. The overall feedback for all devices was positive and did not leave any concerns that completing the survey on the smartphone or tablet would be an impediment to completion.

Device-Specific Difficulties: Fortunately, most participants did not encounter any serious difficulties with any of the devices, though the need for some small areas for improvement did arise. When using the smartphone, some participants indicated that they had difficulty selecting options that were at the bottom of the screen. Participants discovered that this issue could be resolved by scrolling up and then down again before trying to select options at the bottom of the screen. This issue should be tested prior to the full launch of the survey to determine if the problem is universal or if it was related to the specific device used in testing.

Two participants during week 1 had difficulty viewing the entire website unless the tablet was oriented horizontally in “landscape” mode. At other points, participants indicated that they had to “zoom in” with the tablet to select a bubble or write in a response, but this did not present any major issues for any participant.

Survey Confusion

Survey Introduction: The introduction text to the survey was identified as being too lengthy, with participants saying they would not pay attention to it because it was too wordy. This led to some minor changes to the introduction language between rounds 1 and 2 of usability testing. The revised text reads: “Welcome to the 2015 Survey on Consumer Behaviors. The purpose of this survey is to explore consumer’s behavior with respect to specific goods and services in the United States.”

Multiple Record Instructions: One of the greater areas of confusion concerned parts of the instructions added to the survey to clarify how the records would be entered. Specifically, the language that explained to participants in the web-based survey that they would be entering their information one record at a time caused notable difficulty.

> “I have to read it again because I want to be sure of what you’re asking. I think you’re saying if I ate at two different places to record them separately. Like if I ate at Chipotle twice, I should record them both.”
>
> “I wish there was a button where you know you could do a second transaction. At the end, you get the option for the second transaction, which is good and it was good they said in between to record them in the same way.”

- "After a couple of times reading it, I figured it out. Especially for immigrants like me, they should have made it simpler. When you read it, it's easier ... they could make it easier."

Although it appeared that most participants were able to eventually understand the instructions, either by rereading it carefully or by moving through and trying to complete the survey, it was felt that the language was causing too much cognitive burden. We revised the language after the first week of usability testing to try to reduce participant confusion. All services except for the "Taxi, Limousine, Rideshare, or Shuttle Service" category had specific language written to better explain how to record multiple transactions within that service, with appropriate examples for each service category. The "Taxi, Limousine, Rideshare, or Shuttle Service" category did not receive updated language because it was determined that there were no likely scenarios in which a respondent could enter multiple payments for the same cab, so revised language was not necessary.

During the second week of testing, 10 of 12 participants demonstrated the ability to comprehend the instructions to only enter one transaction at a time. Of the two that did not understand at first, one participant was able to figure out as he went along that he was supposed to enter only one transaction at a time. The other participant explicitly said that he was trying to enter transactions at both a fast food establishment and an ice cream parlor, and he only entered one of the two transactions, as he was unable to comprehend how to enter more than one. However, despite leaving one of his transactions out, his data was still valid for the transaction that he was able to enter. Therefore, we decided to leave the survey language as-is, as there is no further revision to be made that would result in more participants taking the time to read all the instructions.

Language and Response Option Findings

<u>Restaurants or Other Prepared Food/Drink Services:</u> No major changes or additions were identified during round 1 of usability testing; participants in this round by and large understood all of the categories and did not think that any categories were missing. However, two participants near the beginning of round 2 expressed confusion over where to put restaurants such as Chipotle Mexican Grill and Panera Bread. One participant said, "I'd put it under fast food, but it's not fast food; it's fast casual." The other said, "I guess you'd call [Chipotle] a cafeteria." Given that two participants said they would classify the same establishment as two different categories, a "Fast Casual" category was added halfway through round 2 of testing for these types of establishments.

<u>Hotel/Motel:</u> There was some minor confusion regarding how to log certain hotel-related tips and expenses. The room fee, for example, is not a "service" in the same manner as housekeeping or room service. This caused one participant to try to record his hotel bill as the "total bill amount" before recording a tip he had left for the "housekeeping" service. Before week 2 of usability testing, the term "payment" was changed to "transaction" to help participants consider all of these financial interactions as service transactions instead of just the direct payments to the establishment. Additionally, participants were asked to record the number of days they stayed and the room fee if they reported that they had stayed in a hotel prior to recording any other services they might have used. This was meant to remove the desire to report the room bill under any service record.

Personal Grooming, Beauty, or Massage Services: This service category required only minor edits overall, because most participants had no major suggestions based on the subcategories that had been added during cognitive testing. Before round 2 of usability testing, a service was added for "Makeup artist" based on the comments of two participants in round 1 who indicated these were professionals who work in makeup stores or individuals who do makeup for groups at events such as weddings. One of these participants reported that she had one such job in the past and had received tips for her service. No additional changes were uncovered during round 2 of usability testing.

Moving or Household Maintenance: No major changes or additions were identified during round 1 of usability testing; participants in this round by and large understood all of the categories and did not think that any categories were missing. Halfway through round 2 of testing, the response options for this section were revised further, as one participant expressed confusion about how to classify ordering and paying for moving boxes. Though this expenditure is clearly a moving expense, none of the four categories presented accounted for this type of expense. Therefore, an additional category called "Equipment Rental" was established midway through round 2 of testing. Although equipment rental is not an expense where people would typically leave a tip, it does fall under the "Moving or Household Maintenance" category, so providing a category for participants to place this type of expense is necessary to ensure the quality of the data is not compromised by participants who may put these expenses improperly into one of the other categories.

Casino: The subcategories for this group of services was somewhat difficult to evaluate because several participants indicated that they did not have enough experience with casinos to be able to speak with confidence to any of the subcategories listed under "Casino." However, the categories that were listed did not confuse participants, even among those who did not have much experience with casinos. That being said, two participants in round 1 felt that another option should be included for valet services, so this option was added before round 2.

Halfway through round 2 of testing, the "Casino" section was revised further after one participant expressed confusion about the transaction amount for the dealer. After selecting that he paid cash for the transaction or service, he was confused about the amount, and asked, "Do I put the amount of money that I lost?" He was unable to figure out the amount of the actual bill because one does not receive a bill for gambling services; one only pays the amount that he or she loses at the table. For this type of transaction, it was decided not to ask whether and how much participants paid for the service. Instead, the survey skips this question for anyone who says that their service was a casino dealer and proceeds directly to the questions regarding automatic and voluntary tips.

Taxi, Limousine, Rideshare, or Shuttle Service: There was some minor confusion expressed concerning the term "Rideshare" in round 1 when discussing services such as Uber and Lyft. One participant stated that he would have thought the survey was referencing carpooling or similar services if the examples for Uber and Lyft were not provided. Another participant felt that Uber would have its own category. When these participants were asked about how the language should be revised and what they would call such services if not "Rideshare," they had no clear suggestions. Given that all participants were able to understand the purposes of that subcategory upon reading, only small revisions were made to the language to emphasize Uber as the primary service. During round 2, language was revised to read "Uber, Lyft, or other Ride-Share service," rather than "Ride-

Share service (e.g., Uber or Lyft)." Participants in round 2 who were asked all said that they understood what the category entailed.

Paying for a Product or Service: Some participants in round 1 indicated that they would sometimes go somewhere and they would not pay any portion of the bill (e.g., a friend or a parent would pay the bill). In the first version of the survey, participants were expected to click "Not Applicable – There was no bill for this service" when asked how they paid for the service. However, this caused confusion, as there oftentimes was still a bill for the service; it was just paid by someone else (i.e., not the person taking the survey). Before round 2 of testing began, a follow-up question was added that read, "Did you pay for this product or service?" If participants answered "yes," they were asked the appropriate follow-ups about the bill payment method and amount. If they answered "no," they were no longer asked about the amount and method of payment of the bill and were instead directly routed to the tipping questions. The response option "Not Applicable – There was no bill for this service" was removed from the payment method screen.

"I don't remember": In round 1 of usability testing, participants had the option to select "I don't remember" if they could not remember the amount of their bill or their tip. One participant selected this option, and upon probing, revealed that she was in a group of people and they each paid their portion of the tip. She could not recall the exact amount, so she selected "I don't remember." After discussion, it was decided that rounded amounts in these instances were preferable to no numerical data at all, so the decision was made to remove the "I don't remember" response option from these questions before round 2 of testing. No participants in round 2 expressed a desire for an "I don't remember" option.

Total Bill, Automatic Tip, and Voluntary Tip Language: For the question regarding the total bill amount, there were clear indications that some participants did not fully read the instructions concerning the total bill amount, and that it was meant to exclude any tips, whether voluntary or automatic.

The question concerning automatic tips caused some minor confusion among round 1 participants. Some indicated that while they knew that some places did this, they were confused by its inclusion in the survey. One participant in round 1 selected the option on the assumption that there was an automatic tip of some kind included in the bill for the shuttle service. This indicated that some respondents might report that there were automatic tips added by the business when they were not sure. The chance for respondents to report that a business was including an automatic tip without knowing if this occurred was concerning, because it could lead to false estimates about how often businesses add such tips. Before round 2 of testing, this response option "Yes, but amount not provided" was dropped from the survey, leaving respondents with the option to either provide an amount of automatic gratuity or report that one was not included.

Non-Monetary Gifts: The question concerning non-monetary gifts and payments continued to cause participants trouble during week 1 of usability testing. Some confusion centered on what a non-monetary gift was and there were indications that examples could be clearer. Specifically, the language for "coupons" as a non-monetary gift was an area of confusion. One participant read that and started to believe it was for reward points that you can get from restaurants and other establishments. Despite this confusion, most were able to understand it if they gave sufficient effort to reading and interpreting the definition, but it was determined that streamlining the language and question flow would still be necessary for round 2 of testing. Therefore, in round 2 of testing, participants were no longer asked, "Did you leave a non-monetary tip?" as a separate question. Instead, "Non-monetary*" was added as a payment option under the question, "What payment type(s) did you use to pay the voluntary tip?" The examples given were also changed from "coupon or event tickets" to "concert tickets, a bottle of wine, or a meal." Explanatory text was added to the bottom of the response option bank that read: "*Examples of non-monetary gifts are: concert tickets, a bottle of wine, or a meal. Note that non-monetary gifts should only be recorded if they were used to compensate for the service. Non-monetary gifts that are given as personal tokens of appreciation should not be recorded." If participants selected this option, they were then asked to estimate the value of the portion of the tip that was non-monetary in a follow-up question.

Another potential flaw with this item was discovered with the round 1 participant who had given non-monetary gifts to her movers. In this case, she gave several gifts and listed them, but when it came to evaluating the worth of the items, she was only able to successfully value the price of the food she had provided. She determined that evaluating the worth of the furniture that she had given away was not possible to do accurately and did not include those items in her evaluation. This issue serves as evidence that this item will lead to some problematic data in the survey, but there is no clear manner in which it can be fixed, because evaluating the dollar value of used items is cognitively challenging.

Ultimately, while there were some signs that respondents were able to grasp the concept of the question, there were numerous signs of respondent confusion despite multiple revisions. Although this information about non-monetary tips could be useful for providing estimates of income that is not currently captured by the other survey measures, there are numerous opportunities for error that could enter such records. Serious discussion about whether this item is necessary to the survey effort will be required prior to the pilot test. Respondent error, burden, and low incidence rates could make the inclusion of the item more problematic than the benefits of this information warrant.

<u>Recording Non-Tipping Occasions:</u> Some participants indicated that they wanted to list payments for services that fell within the realm of some of the major service categories but were not services that could be considered tipping events. Participants in some instances indicated that they thought they should be able to record transactions such as buying groceries or purchases made at the pharmacy. Another participant, when looking at the options for the "Transportation" category, thought that there should be subcategories for public transit options such as buses and metro. A different participant recorded a transaction under the "Personal Grooming, Beauty, or Massage Services" category for a few dollars. When asked about this, he indicated that the purchase was for hair gel and that the grooming category should be clear that it was not meant for store-bought purchases.

Other Findings

<u>Participant Tipping Experience:</u> Seven of the 12 participants in round 1 and 6 of 11 participants in round 2 indicated that they had worked for a job at some point that received tips for service. Of those, nearly all of them (five of seven in round 1, four of six in round 2) said that experience influenced their tipping behaviors, either by increasing their knowledge about how important tips are

to some professions or from an increased sense of empathy for people who work those professions because of their shared experience.

Participant Perceptions of Survey Focus: Perceptions of the survey intent were mixed. Some participants felt that the aim of the survey was to better understand consumer spending patterns and what people were spending their money on. Two participants thought that the survey might be used to create changes in legislation affecting the minimum wage of certain tipped jobs. Four of 12 round 1 participants and two of 10 round 2 participants seemed to have a good sense of the purpose of the study. Of those six participants, one said that knowing the purpose of the study would negatively affect how truthful he would be when filling out the survey.

Banking: Across both rounds of usability testing, there was no consensus concerning the use of banking records to help complete the survey, although nearly all reported that they had a banking profile or records they could check. Seven participants in round 1 and four in round 2 indicated that they had a banking profile and might look up their records to complete the survey, while five in round 1 and two in round 2 indicated that they would not look up bank records or receipts. A few participants indicated that they would be more likely to use a banking profile to check their records for longer periods of recall, such as for the 5-day condition.

Summary and Conclusions

Survey Variant Selection: Although numerous changes came from the testing process, the first and most pressing issue was to determine which of the two proposed survey versions would be the better option. Ten of the 12 participants indicated that they preferred thinking about each service separately and felt that it was easier for them to complete the second version. Although the increased survey burden is a concern, the benefits clearly outweighed the costs given the preference, leading to the decision to move forward with the second version for usability testing.

Online Survey Design: The next major design decision was about the setup for the web version of the survey. Two drafts were created using two different survey platforms. The Verint version showed a greater amount of questions on the screen at once but was ultimately deemed not intuitive enough to complete on mobile devices. The Confirmit platform required respondents to answer one question at a time but was much more intuitive to complete on mobile devices and was thus selected for use. However, the negative consequence of its selection was an increase in survey burden due to the format of answering one question at a time. However, final specifications of the pilot survey will depend on the survey platforms operated by the population vendors, Ipsos and GfK. FMG will work in consultation with both to ensure that there is common programming specifications and that the final product aims to minimize burden across devices.

Device Preference: One of the main concerns in the survey design stage during usability testing was ensuring that the prototype was compatible not only with computers but also with smartphones and tablets. The thought was that many survey respondents would complete the survey with the device that was most accessible and that in numerous situations that would be a smartphone or other mobile device. Respondents generally indicated some preference for the laptop and phone versions of the survey, and most reported that they would complete the survey on such a device, either because of preference or accessibility reasons. Fortunately, no respondent felt that the survey was much more difficult to complete on a smartphone than on the other devices, and a few commented on how it seemed that the survey was well designed for such platforms.

Recall and Accuracy: Although respondents seemed to indicate that a 3-day recall period would be the maximum amount of time that they could accurately record their expenditures, it was determined that the pilot study should proceed with a 1-day recall period rather than expanding to a 3-day period because the maximum recall that could be tested would likely increase difficulty significantly. This determination was also made because of concerns about survey burden and from a lack of irrefutable proof that respondents could accurately record expenditures from multiple days earlier. Although there were respondents who indicated that they might check their banking records for greater accuracy, this was not observed in the lab and cannot be expected during the survey fielding. Although some respondents will not have any records because of the shorter recall time, this is considered an acceptable consequence given the importance of maximizing the accuracy in the records that are gathered.

Wording Changes and Service Additions: Numerous questions from the original draft of the survey required revision to enhance respondent comprehension and include directions specific for the web-based version. Most items received revisions at some point of the cognitive or usability testing process, but no major items were added or removed throughout beyond the addition of instructions or additional service categories. The language edits did not change the original intent of the questions in any manner. The primary revisions were the instructions logging multiple entries, revised and new service categories, and clarification to the definition of “non-monetary” gift.

Other Findings: Other noteworthy findings from usability testing included participant thoughts on the use of banking statements or other records to accurately fill out the survey and the general intent of the survey. Most participants reported that they had access to some kind of online banking profile or other means of tracking their expenditures, and roughly half of those felt that they might check those records to complete such an online survey. However, no participants appeared to actually check these records while in the lab setting, so it is difficult to gauge how often this might occur during the pilot test, if at all. Finally, a notable portion of participants were able to successfully guess at the general intent of the survey (i.e., that it was an IRS effort aimed at better determining the tipped income that might not be reported), but nearly all claimed this knowledge would not negatively influence whether they would accurately fill out such a questionnaire (one respondent acknowledged that this might negatively influence how truthfully he filled it out). This feeling was endorsed by others when they were told the true purpose of the study. A few participants acknowledged that others might not be inclined to fill out the survey properly if they knew the purpose.

Outstanding Issues to be Resolved Prior to Pilot Study: The two major issues that remain prior to launching the pilot survey are the questions about non-monetary gifts and survey platform design. Given the predicted low-incidence rate and high degree of confusion presented by the questions and language concerning non-monetary gifts, it is FMG’s recommendation that these questions be removed from the survey. Even after multiple attempts to clarify the language there were still difficulties in interpreting this type of gift, in addition to other issues such as properly valuing gifts. The survey design issues will be addressed in coordination with the panel vendors, Ipsos and GfK, as they will likely both separately program the final survey. For this reason, it will be important to give both detailed specifications about how the survey should be designed and programmed before attempting to resolve any discrepancies between their unique systems. Respondent burden and mobile accessibility are the two greatest design elements that need to be addressed during this process.

Consumer Tipping Draft Survey for Usability Testing

Welcome to the 2015 Survey on Consumer Tipping Behaviors. This survey is aimed at determining average expenditures amongst consumers, particularly tipping related expenditures. In this short survey, we will ask you about your expenditures within the past XX days. This survey is being conducted by a third party research group, Fors Marsh Group, LLC.

This survey should only take 8 minutes to complete.

Screener1) In the past XX days, have you made any expenditures at a restaurant, casino, hair stylist, hotel/motel, taxi/limousine service, or on a cruise ship.

1A) Please provide details for any expenditures made in the past <day/week/etc.> at a restaurant, casino, hair stylist, hotel/motel, taxi/limousine service, on a cruise ship, or at an auto mechanic. If you have made multiple expenditures on a given type of service in the past <day/week/etc.>, provide separate details for each. If you have not made any expenditures on one of the listed services in the past <day/week/etc.>, select "No Expenditure".

Establishment/Service Type (restaurant, casino, hair stylist, hotel/motel, taxi/limo, cruise ship)	*Sub-Type (e.g., for restaurants: Café/Family-Style/Diners, Traditional Restaurants/Casual Dining, Upscale Casual Dining, Fine Dining, Fast Food, Delivery, Ice Cream, Coffee Shops, Smoothie, Self-Service/Cafeteria/Buffets)*	*Total bill amount (after tax, before automatic or voluntary gratuity)*	*Payment type for bill (cash, debit, credit, check, gift card, smartphone credit or app, paper or online coupon {e.g., Groupon}, non-monetary, other)*	*Amount of automatic gratuity added by establishment*	*Amount of voluntary tip*	*Payment type for Voluntary tip (same options as column 4)*	*Amount of non-monetary gift**	*Description of tip if non-monetary (text field)*
Drop down menu	Drop down menu	Text	Multiple choice (select all that apply)	Text	Text	Multiple choice (select all that apply)	Text	Text

**If a portion of the gratuity or tip took the form of a non-monetary payment (e.g., a coupon or event tickets) indicate the cash equivalent amount in this column. Note that non-monetary transfers should only be counted as tips if they were used as payment for immediate service and are used as a substitute for a monetary tip. Non-monetary transfers that are used as gifts/personal tokens of appreciations should not be counted as tips.*

[Note: 1B is an alternative question format that could be tested during the usability phase. This method would ask a variant of question 1 for each of the services and establishments of interest. The goal for this approach is to improve participant recall and have them actively consider each type of establishment we are interested with. If they answer yes, they get a follow-up question asking them to list their expenditures for that type of establishment]

1B) In the last <day/week/etc.>, have you purchased/visited a <list each *Establishment/Service Type (restaurant, casino, hair stylist, hotel/motel, taxi/limo, cruise ship, auto mechanic)*>?

1. No
2. Yes

[If Q1 is yes, list the table below for the service from the prior question]

Please answer the following questions regarding the amount spent and the amount tipped on this purchase/visit.

Sub-Type (e.g., for restaurants: Café/Family-Style/Diners, Traditional Restaurants/Casual Dining, Upscale Casual Dining, Fine Dining, Fast Food, Delivery, Ice Cream, Coffee Shops, Smoothie, Self-Service/Cafeteria/Buffets)	*Total bill amount (after tax, before automatic or voluntary gratuity)*	*Payment type for bill (cash, debit, credit, check, gift card, smartphone credit or app, paper or online coupon {e.g., Groupon}, non-monetary, other)*	*Amount of automatic gratuity added by establishment*	*Amount of voluntary tip*	*Payment type for voluntary tip (same options as column 4)**	*Amount of non-monetary gift**	*Description of tip if non-monetary (text field)*
Drop down menu	Text	Multiple choice (select all that apply)	Text		Multiple choice (select all that apply)	Text	Text

**If a portion of the gratuity or tip took the form of a non-monetary payment (e.g., a coupon or event tickets) indicate the cash equivalent amount in this column. Note that non-monetary transfers should only be counted as tips if they were used as payment for immediate service and are used as a substitute for a monetary tip. Non-monetary transfers that are used as gifts/personal tokens of appreciations should not be counted as tips.*

[Note: Demographic items 2-4 will be captured by the frame file of both survey panels and will not be asked of participants in the actual survey.]

2) What is your age?

   <Text box>

3) In which <county/ZIP code> do you live?

   <Drop-down menu>

4) What is your gender?

   1. Male
   2. Female

5) Are you of Hispanic or Latino origin (ethnicity)?

   1. Yes, of Hispanic origin
   2. No, not of Hispanic origin

6) What is your race? Please select one or more. Are you...

   1. White
   2. Black or African-American
   3. Asian
   4. Native Hawaiian or Other Pacific Islander
   5. American Indian or Alaskan Native

7) Please indicate your highest level of educational attainment:

   1. No formal education
   2. 1st, 2nd, 3rd, or 4th grade
   3. 5th or 6th grade
   4. 7th or 8th grade
   5. 9th grade
   6. 10th grade
   7. 11th grade

8. 12th grade NO DIPLOMA
9. HIGH SCHOOL GRADUATE - high school DIPLOMA or the equivalent GED)
10. Some college, no degree
11. Associate degree
12. Bachelors of degree
13. Master's degree
14. Professional or Doctorate degree

8) Please indicate your employment status:
   1. Working - as a paid employee
   2. Working - self-employed
   3. Not working - on temporary layoff from a job
   4. Not working - looking for work
   5. Not working – retired
   6. Not working – disabled
   7. Not working – other

9) Please indicate your annual household income:

   1. Less than $5,000
   2. $5,000 to $7,499
   3. $7,500 to $9,999
   4. $10,000 to $12,499
   5. $12,500 to $14,999
   6. $15,000 to $19,999
   7. $20,000 to $24,999
   8. $25,000 to $29,999
   9. $30,000 to $34,999
   10. $35,000 to $39,999
   11. $40,000 to $49,999
   12. $50,000 to $59,999
   13. $60,000 to $74,999
   14. $75,000 to $84,999
   15. $85,000 to $99,999
   16. $100,000 to $124,999
   17. $125,000 to $149,999
   18. $150,000 to $174,999
   19. $175,000 or more

The Paperwork Reduction Act requires that the IRS display an OMB control number on all public information requests. The OMB Control Number for this survey is 1545-1349. We estimate the time required to be eight minutes. Also, if you have any comments regarding the time estimates associated with this study or suggestions on making this process simpler, please write to:

Internal Revenue Service
Tax Product Coordinating Committee

1111 Constitution Avenue NW
Washington, DC 20224

Appendix B: Survey Edits from Cognitive Testing

- Question 1 originally read: “Please provide details for any expenditures or payments made in the past week at a restaurant or other prepared food service, casino, barber or hair stylist, hotel/motel, or a taxi/limousine service. If you have made multiple expenditures on a given type of service in the past week, provide separate details for each. For example, if you stayed at a hotel and had tipped room service and a concierge, please record those separately.” Question 1 was changed to: “Please provide details for any expenditures or payments made in the past week at a restaurant or other prepared food/drink service, casino, personal grooming or beauty service, moving or household cleaning/maintenance, hotel/motel, or a taxi/limousine/shuttle service. If you have made multiple expenditures (e.g., bill, tip) at a given establishment or type of service in the past week, provide separate details for each. For example, if you stayed at a hotel and had tipped room service and a concierge, please record those separately.”
- Question 2A originally read: “In the last week, have you made any expenditures at a restaurant or other prepared food service?” This question was changed to: “In the past week, have you made any expenditures at a restaurant or other prepared food/drink service?”
- For Question 2A, the following response options were added to the “Restaurant or Other Prepared Food/Drink Service” category:
    - “Bar”
    - “Coffee Shops”
    - “Food Cart/Truck”
    - “Self-Service/Cafeteria/Buffets”
- For Question 2B, the following response options were added to the “Casino” category:
    - “Bar”
    - “Full-Service Dining”
    - “Self-Service/Cafeteria/Buffets”
- For Question 2C, the question originally read: “In the last week, have you visited a barber or hair stylist?” This question was changed to: “In the last week, have you made any expenditures on *personal grooming, beauty, and massage services?”*
- For Question 2C, the following response options were added to the new “Personal Grooming, Beauty, and Massage Services” category:
    - “Manicurist/Pedicurist”
    - “Massage Therapist”
    - “Waxing/Hair Removal”
    - “Facial/Skin Care”
- For Question 2C, there originally were no instructions informing participants to record transactions separately. These instructions were added: “Provide separate details for each expenditure. For example, if you had tipped a hair stylist in addition to tipping a manicurist, please record those separately.”
- Question 2D was added to the survey: “In the last week, have you made any expenditures on moving or other household maintenance services?” The following response options were added to the “Moving or Household Maintenance” category:
    - “Professional Movers”
    - “Maid or Cleaning Service”
    - “Lawn/Gardening service”

- "Handyman/ Repairman"
- For Question 2E, the following response options were added to the "Hotel/ Motel" category:
  - "Bar"
  - "Full-Service Dining"
  - "Self-Service/ Cafeteria/ Buffets"
- For Question 2F, the question originally read: "In the last week, have you used a taxi or limousine service?" This question was changed to: "In the last week, have you used a taxi, limousine, rideshare, or shuttle service?"
- For Question 2F, the response option "App-Based Taxi" was changed to "Ride-Share service (e.g., Uber or Lyft)"
- First non-monetary gift question changed from "Value of non-monetary gift* you provided" to "Did you give a non-monetary gift* for this service? If so, can you estimate its value?"
- Non-monetary gift description was changed from the following: "*If a portion of the gratuity or tip took the form of a non-monetary payment (e.g., a coupon or event tickets), indicate the cash equivalent amount in this column. Note that non-monetary transfers should only be counted as tips if they were used as payment for immediate service and are used as a substitute for a monetary tip. Non-monetary transfers that are used as gifts/ personal tokens of appreciation should not be counted as tips." The revised description: "*A non-monetary gift could be something like a coupon or event tickets. Note that non-monetary gifts should only be recorded if they were used to compensate for service. Non-monetary gifts that are given as personal tokens of appreciation should not be recorded."

Appendix C: Survey Edits from Usability Testing

- Changed introduction from the following: "Welcome to the 2015 Survey on Consumer Tipping Behaviors. The purpose of this survey is to determine payments for commonly tipped services in the United States." The revised introduction: "Welcome to the 2015 Survey on Consumer Behaviors. The purpose of this survey is to explore consumer's behavior with respect to specific goods and services in the United States."
- Changed "In the last [TIME], have you made any service-related payments at a..." to "In the last [TIME], have you made any transactions at a..." for each service category question.
- Changed the uniform language for multiple payments for most services to a unique question for each with examples. Text for "limousine, taxi, or shuttle service" service category was not altered because it was determined that it was not reasonable to expect that someone could make multiple payments to the same cab. Language changed from the following: "On the next page, we will ask you to record one transaction you have made for [SERVICE CATEGORY]. Do not record transactions for which you have already provided information."
  - "On the next page, we will ask you to record one transaction you had at a restaurant or other prepared food/drink service. If you have had multiple transactions at the same establishment (even if during the same visit), please record each transaction separately. For example, if you made separate payments for a drink at the bar and a meal at the table, please record these transactions separately."
  - "On the next page, we will ask you to record one transaction you have made at a hotel/motel. If you have engaged in multiple transactions at the same establishment (even if during the same visit), please record each transaction separately. For example, if you engaged in separate transactions for valet service and luggage assistance during the same visit, please record these transactions separately. Do not record transactions for which you have already provided information."
  - "On the next page, we will ask you to record one payment you have made at a personal grooming, beauty, or massage service. If you have made multiple payments at the same establishment (even if during the same visit), please record each transaction separately. For example, if you made separate payments to your hair stylist and your manicurist during the same visit, please record these transactions separately. Do not record transactions for which you have already provided information."
  - "On the next page, we will ask you to record one payment you have made at a moving or household maintenance service. If you have made multiple payments at the same establishment (even if during the same visit), please record each transaction separately. For example, if you made separate payments to your gardener and your landscaper during the same visit, please record these transactions separately. Do not record transactions for which you have already provided information."
  - "On the next page, we will ask you to record one transaction you have made at a casino. If you have engaged in multiple transactions at the same establishment (even if during the same visit), please record each transaction separately. For example, if you engaged in separate transactions to your casino dealer and your floor server

while playing at the same table, please record these transactions separately. Do not record transactions for which you have already provided information."

- Added "Fast Casual" as a service for the "Restaurant or Other Prepared Food/Drink Service" category.
- Added questions asking "What was the average nightly rate for the room?" and "How many nights did you stay at this hotel?" after indicating that they had a transaction at a hotel/motel.
- Added "Makeup Artist" as a service for the "Personal Grooming, Beauty, or Massage Services" category.
- Added "Equipment Rental" as a service for the "Moving or Household Maintenance Services" category.
- Added "Valet" as a service for the "Casino" category.
- Changed "Ride-Share service (e.g., Uber or Lyft)" to "Uber, Lyft, or other Ride-Share service."
- Added question after they select their service for "Did you pay for this product or service? (Yes/No)." If "yes," they move to the payment options for the bill. If "no," they move to the question asking if an automatic tip was added by the business.
- Removed response option "Not Applicable – there was no bill for this service." from the question asking what payment type they used to pay their portion of the bill.
- Bolded the language "(after tax, before automatic or voluntary tip)" for the question about the amount of the bill paid.
- Removed the response option for the automatic tip question stating, "Yes, but amount not provided."
- Added question "Did you leave a voluntary tip for this service?" after the question for the automatic tip. If "yes," they move forward to the question about the type of tip.
- Removed the response option for the voluntary tip type question "There was no tip for this service."
- Included description of non-monetary gift for the question of voluntary tip payment type. Removed description from follow-up questions about non-monetary tip value. "*Examples of non-monetary gifts are: concert tickets, a bottle of wine, or a meal. Note that non-monetary gifts should only be recorded if they were used to compensate for the service. Non-monetary gifts that are given as personal tokens of appreciation should not be recorded."
- Follow-up non-monetary gift question changed from "Did you give a non-monetary gift* for this service? If so, can you estimate its value?" to "Estimate the value of the part of the tip that is non-monetary."

2017 IRS Tipping Questionnaire
[Modified December 23, 2016]

Revisions made on December 12 are highlighted in pink.

Changes compared with the 2015 version of this questionnaire:

- Updated standard screeners to the most recent versions.
- Changed 2015 to 2016 in the intro on page 6.
- Changed 2016 to 2017 in March 2017 during fielding.
- Updated the OMB number on the last page to "1545-2261".
- Replaced prior programming notes referring to "DEM_4" with "[SKIP TO USRETH3]".
- Moved race & ethnicity question towards the end of the questionnaire.

Definition of 1,500 completes (tippers) per month:

- Answer "01" (Yes) to any of Q1_A, Q2_A, through Q6_A questions <u>AND</u>
- Also answer "01" (Yes) to corresponding _G questions.

Questionnaire and Programming Notes (PNs):

RECORD START AND STOP TIMESTAMPS FOR EACH RESPONDENT SESSION.

RE-ENTRY PROHIBITED "AS SOON AS "YEAR / MONTH" IS ANSWERED. IF A RESPONDENT LEAVES THE SURVEY, DO NOT ALLOW HIM/HER TO RE-ENTER."

IF THEY ENTER THE SURVEY AND LEAVE WITHOUT COMPLETING THE YEAR/MONTH QUESTION, AND THEN RE-ENTER – THE DATE AND TIME VARIABLE SHOULD BE FOR RE-ENTRY."

ALL QUESTIONS EXCEPT FOR CORTEX STANDARD SCREENERS ARE NON-MANDATORY: IF RESPONDENT DOES NOT ANSWER QUESTIONS Q1_B, B2_B, Q3_B, Q4_B, Q5_B, AND Q6_B, THE RESPONDENT SHOULD STAY ON THE SAME PAGE AND BE SHOWN THE "MISSING ANSWER(S)" VALIDATION, AFTER WHICH THEY SHOULD BE ALLOWED TO MOVE TO THE NEXT QUESTION. AND AUTO-CODE AS REFUSED
[DO NOT SHOW THE "REFUSED" CODE TO RESPONDENTS. IF A RESPONDENT CHOOSES NOT TO ANSWER A QUESTION (THEY HIT "NEXT" BUTTON WITHOUT ANSWERING ANYTHING) AUTOCODE THEM AS "REFUSED" AND FOLLOW ANY LOGIC THAT APPEARS AFTER THE "REFUSED" CODE ON EACH QUESTION.

[Cortex 5 Standard Screener: DO NOT MODIFY]

YEAR/MONTH. What is your date of birth?

- YEAR
- _1910 1910
- ...
- _2015 2015

[Cortex 5 Standard Screener: DO NOT MODIFY]

RESP_AGE. Hidden Question - RESP_AGE "this is a dummy question that will hold age"

- ☐ USE RESP_AGE response list

[PN] 18+ only

[Cortex 5 Standard Screener: DO NOT MODIFY]

QUOTAGERANGE. Hidden Question - QUOTAGERANGE "this is a dummy question that will hold age breaks" for the quotas that should be defined by the PM; it CAN be edited and lines can be added to meet survey objectives.

[PN] Define age groups as follows:

- ☐ 18 to 24
- ☐ 25 to 29
- ☐ 30 to 34
- ☐ 35 to 39
- ☐ 40 to 44
- ☐ 45 to 49
- ☐ 50 to 54

- ☐ 55 to 59
- ☐ 60 to 64
- ☐ 65 to 69
- ☐ 70 to 74
- ☐ 75+

[Cortex 5 Standard Screener: DO NOT MODIFY]

RESP_GENDER. What is your gender?

- ☐ _1 Male
- ☐ _2 Female

[Cortex 5 Standard Screener: DO NOT MODIFY]

[PN: Hidden question for internal panel (collected from CI variable). 5-7 responses will be shown to respondent; it must include qualifying countries & _999 OTHER. List determined by CS or SW]

Country10. In which country do you live?

- ☐ USE Country10 response list
    - Australia
    - Canada
    - México
    - United States
    - United Kingdom
    - Other

[PN] Only US may proceed. Terminate other respondents.

[Cortex 5 Standard Screener: DO NOT MODIFY]

[PN: Add term instruction if invalid zips are not allowed; no GEO info will be populated from invalid zips]

QMktSize_US. REQUIRED Please insert your zipcode:

[PN] Terminate if zipcode is invalid.

[Cortex 5 Standard Screener: DO NOT MODIFY]

HCAL_REGION1_Label_US. Hidden Question: State

- ☐ USE HCAL_REGION 1_Label_US response list

[Cortex 5 Standard Screener: DO NOT MODIFY]

HCAL_Region2_Label_US. Hidden Question: DMA

- ☐ USE HCAL_Region2_Label_US response list

[Cortex 5 Standard Screener: DO NOT MODIFY]

HCAL_STDREGION_4CODES_Label_US. Hidden Question: Census Region

- ☐ (1) Northeast
- ☐ (2) Midwest
- ☐ (3) South

- ☐ (4) West

[Cortex 5 Standard Screener: DO NOT MODIFY]

HCAL_STDREGION_Label_US. Hidden Question: Census Division

- ☐ (1) New England
- ☐ (2) Middle Atlantic
- ☐ (3) East North Central
- ☐ (4) West North Central
- ☐ (5) South Atlantic
- ☐ (6) East South Central
- ☐ (7) West South Central
- ☐ (8) Mountain
- ☐ (9) Pacific

[Cortex 5 Standard Screener: DO NOT MODIFY]

Time Zone_Label_US. Hidden Question: Time Zone

- ☐ (5) Eastern (GMT -05:00)
- ☐ (6) Central (GMT -06:00)
- ☐ (7) Mountain (GMT -07:00)
- ☐ (8) Pacific (GMT -08:00)
- ☐ (9) Alaska (GMT -09:00)
- ☐ (10) Hawaii-Aleutian Islands (GMT -10:00)

[Cortex 5 Standard Screener: DO NOT MODIFY]

[PN: USRETH3 can be asked alone (without USRACE4)]

[Cortex 5 Standard Screener: DO NOT MODIFY]

USEDU3. What is the highest degree or level of school you have completed?

Select only one

- ☐ Education through Grade 12 [Expandable Header]
  - _1 Grade 4 or less
  - _2 Grade 5 to 8
  - _3 Grade 9 to 11
  - _4 Grade 12 (no diploma)
- ☐ High School Graduate [Expandable Header]
  - _5 Regular High School Diploma
  - _6 GED or alternative credential
- ☐ College or Some College [Expandable Header]
  - _7 Some college credit, but less than 1 year
  - _8 1 or more years of college credit, no degree
  - _9 Associate's degree (AA, AS, etc)
  - _10 Bachelor's degree (BA, BS, etc.)
- ☐ After Bachelor's Degree [Expandable Header]
  - _11 Master's degree (MA, MS, MBA, etc.)
  - _12 Professional degree (MD, DDS, JD, etc.)
  - _13 Doctorate degree (PhD, EdD, etc.)

[Cortex 5 Standard Screener: DO NOT MODIFY]

EMP01. What is your current employment status?

Select only one

- ☐ _1 Employed full-time
- ☐ _2 Employed part-time
- ☐ _3 Self employed
- ☐ _4 Unemployed but looking for a job
- ☐ _5 Unemployed and not looking for a job/ Long-term sick or disabled
- ☐ _6 Full-time parent, homemaker
- ☐ _7 Retired
- ☐ _8 Student/ Pupil
- ☐ _9 Military
- ☐ _10 Prefer not to answer

[Cortex 5 Standard Screener: DO NOT MODIFY]

USHHI3. Please indicate your annual household income before taxes.

- USE USHHI3 response list
  - Less than $5,000
  - $5,000-$9,999
  - $10,000-$14,999
  - $15,000-$19,999
  - $20,000-$24,999
  - $25,000-$29,999
  - $30,000-$34,999
  - $35,000-$39,999
  - $40,000-$44,999
  - $45,000-$49,999
  - $50,000-$54,999
  - $55,000-$59,999
  - $60,000-$64,999
  - $65,000-$69,999
  - $70,000-$74,999
  - $75,000-$79,999
  - $80,000-$84,999
  - $85,000-$89,999
  - $90,000-$94,999
  - [illegible]

MQB

(Main Questionnaire Begins)

PROG: ALL QUESTIONS , ARE NON-MANDATORY . PLEASE REFER TO NOTES ON THE FIRST PAGE OF THIS DOCUMENT FOR MORE DETAILED INFORMATION.

Consumer Tipping Survey

Welcome to the 2016 Survey on Consumer Behaviors. The purpose of this survey is to explore consumer's behavior with respect to specific goods and services in the United States. In this short survey, we will ask you about what, if any, transactions of these types have occurred within the last calendar day. This survey is being conducted by a third party research group, Fors Marsh Group, LLC.

This survey should take 8 minutes or less to complete.

SINGLE PUNCH ANSWER
Q1_A. In the last calendar day, have you made any transactions at a *restaurant or other prepared food/drink service*?

00 No [SKIP TO Q2_A]
01 Yes
-99 Refused [SKIP TO Q2_A]

Instruction Page
On the next page, we will ask you to record one *restaurant or other prepared food/drink service* transaction that you made in the last calendar day. You will have an opportunity to record a separate transaction of this type later. If you had more than one transaction of this type (even at the same establishment and/or during the same visit), please record each separately. Do not record transactions for which you have already provided information. [NEXT]

SINGLE PUNCH ANSWER
Q1_B. What type of service did you receive? Record each transaction separately.

01 Full-Service Dining (e.g., traditional restaurants)
02 Fast Casual
03 Fast Food
04 Carry-out/Delivery
05 Bar
06 Coffee Shops
07 Ice Cream/Smoothie Shops
08 Self-Service/Cafeteria/Buffets
09 Food Cart/Truck
-99 Refused
-100 Valid Skip

///SOFT PROMPT///
IF RESPONDENT REFUSES Q1_B, STAY ON THE SAME PAGE AND WRITE "MISSING ANSWER(S)". IF THEY REFUSE AGAIN THEY SHOULD BE ALLOWED TO CONTINUE

SINGLE PUNCH ANSWER
Q1_C. Did you pay for this particular service (excluding any automatic or voluntary tip)?

00 No [SKIP TO Q1_F]
01 Yes
-99 Refused [SKIP TO Q1_F]
-100 Valid Skip

MULTIPLE PUNCH ANSWER
Q1_D. What payment type(s) did you use to pay your portion of the bill?
(select all that apply)

01 Cash
02 Debit
03 Credit
04 Check
05 Gift Card
06 Smartphone credit or app
07 Paper or online coupon (e.g., Groupon)
08 Other
-99 Refused
-100 Valid Skip

OPEN-ENDED ANSWER
Q1_E. What was the amount of the bill that you paid? When filling in cents please enter a value from 00-99.
(after tax, before automatic or voluntary tip)

$[TEXT BOX].[TEXT BOX] [PROG- RANGE IS $0.01-1,000,000]
-99 Refused
-100 Valid Skip

SINGLE PUNCH ANSWER
Q1_F. Did the business add an automatic tip for this service? If so, how much did you pay? When filling in cents please enter a value from 00-99.

00 No
01 Yes, and the amount was: $[TEXT BOX].[TEXT BOX]
-99 Refused
-100 Valid Skip

SINGLE PUNCH ANSWER
Q1_G. Did you leave a voluntary tip for this transaction?

00 No [SKIP TO Q1_J]

01 Yes
-99 Refused [SKIP TO Q1_J]
-100 Valid Skip

MULTIPLE PUNCH ANSWER
Q1_H. What payment type(s) did you use to pay the voluntary tip?
(select all that apply)

01 Cash
02 Debit
03 Credit
04 Check
05 Gift Card
06 Smartphone credit or app
07 Paper or online coupon (e.g., Groupon)
08 Non-monetary*
09 Other
-99 Refused
-100 Valid Skip

[Instructions at the bottom of response option list] *Examples of non-monetary gifts are: concert tickets, a bottle of wine, or a meal. Note that non-monetary gifts should only be recorded if they were used to compensate for the service. Non-monetary gifts that are given as personal tokens of appreciation should not be recorded.

OPEN-ENDED ANSWER
Q1_I. What was the amount of voluntary tip you paid? When filling in cents please enter a value from 00-99.

$[TEXT BOX].[TEXT BOX] [PROG- RANGE IS $0.01-1,000,000]
-99 Refused
-100 Valid Skip

SINGLE PUNCH ANSWER
Q1_J. Have you made any other transactions at a *restaurant or other prepared food/drink service* in the last calendar day?

00 No [SKIP TO Q2_A]
01 Yes [PROCEED TO NEXT INSTRUCTION PAGE]
-99 Refused [SKIP TO Q2_A]
-100 Valid Skip

Instruction Page
Please record your next transaction in the same way as before. [PROCEED to new record for Q1_B]

[PROG:MAX ITERATION FOR EACH SERIES OF QUESTIONS IS 20: WILL NOT SKIP TO NEXT ITERATION IF WE ALREADY COLLECTED 20 RESPONSES. INSTEAD, THEY SHOULD GO TO NEXT SECTION IN RANDOMIZATION].

///RANDOMIZE DETERMINE NEXT SERIES OF QUESTIONS, SELECT FROM Q2_A – Q6_A. RANDOMLY SELECT AFTER EACH SERIES IS COMPLETE///

[PN: RECORD RANDOMIZATION ORDER]

SINGLE PUNCH ANSWER
Q2_A. In the last calendar day, have you had any transactions at a *hotel/motel*?

00 No [SKIP TO Q3_A]
01 Yes
-99 Refused [SKIP TO Q3_A]

[SHOW Q2_RATE and Q2_NIGHTS on same page]

OPEN-ENDED ANSWER
Q2_RATE. What was the average nightly rate for the room? When filling in cents please enter a value from 00-99.

$[TEXT BOX].[TEXT BOX] [PROG- RANGE IS $0.01-1,000,000]
-99 Refused
-100 Valid Skip

[PN – IF RESPONSE IS NOT NUMERIC, PROMPT "Answer [INSERT RESPONSE] is not numeric"]

OPEN-ENDED ANSWER
Q2_NIGHTS. How many nights did you stay at this hotel?

[TEXT BOX] [PROG- RANGE IS 1-365]
-99 Refused
-100 Valid Skip

[PN – IF RESPONSE IS OUT OF RANGE, PROMPT "Answer must be between 1 and 365 days]

Instruction Page
On the next page, we will ask you to record one *hotel/motel* transaction that you made in the last calendar day. You will have an opportunity to record a separate transaction of this type later. If you had more than one transaction of this type (even at the same establishment and/or during the same visit), please record each separately. Do not record transactions for which you have already provided information.

[NEXT]

SINGLE PUNCH ANSWER
Q2_B. What type of service did you receive? Record each transaction separately.

01 Concierge/Front Desk Staff

02 Housekeeping
03 Room Service
04 Valet
05 Bellhop/ Luggage Assistance
06 Bar
07 Full-Service Dining (e.g., traditional restaurant)
08 Self-Service/ Cafeteria/ Buffets
09 Shuttle Service to/ from Hotel/ Motel
-99 Refused
-100 Valid Skip

///SOFT PROMPT///
IF RESPONDENT REFUSES Q2_B, STAY ON THE SAME PAGE AND WRITE "MISSING ANSWER(S)". IF THEY REFUSE AGAIN THEY SHOULD BE ALLOWED TO CONTINUE

SINGLE PUNCH ANSWER
Q2_C. Did you pay for this particular service (excluding any automatic or voluntary tip)?

00 No [SKIP TO Q2_F]
01 Yes
-99 Refused [SKIP TO Q2_F]
-100 Valid Skip

MULTIPLE PUNCH ANSWER
Q2_D. What payment type(s) did you use to pay your portion of the bill?
(select all that apply)

01 Cash
02 Debit
03 Credit
04 Check
05 Gift Card
06 Smartphone credit or app
07 Paper or online coupon (e.g., Groupon)
08 Other
-99 Refused
-100 Valid Skip

OPEN-ENDED ANSWER
Q2_E. What was the amount of the bill that you paid? When filling in cents please enter a value from 00-99.
(after tax, before automatic or voluntary tip)

$[TEXT BOX].[TEXT BOX] [PROG- RANGE IS $0.01-1,000,000]
-99 Refused
-100 Valid Skip

SINGLE PUNCH ANSWER

Q2_F. Did the business add an automatic tip for this service? If so, how much did you pay? When filling in cents please enter a value from 00-99.

00	No
01	Yes, and the amount was: $[TEXT BOX].[TEXT BOX]
-99	Refused
-100	Valid Skip

SINGLE PUNCH ANSWER
Q2_G. Did you leave a voluntary tip for this transaction?

00	No [SKIP TO Q2_J]
01	Yes
-99	Refused [SKIP TO Q2_J]
-100	Valid Skip

MULTIPLE PUNCH ANSWER
Q2_H. What payment type(s) did you use to pay the voluntary tip?
(select all that apply)

01	Cash
02	Debit
03	Credit
04	Check
05	Gift Card
06	Smartphone credit or app
07	Paper or online coupon (e.g., Groupon)
08	Non-monetary*
09	Other
-99	Refused
-100	Valid Skip

[Instructions at the bottom of response option list] *Examples of non-monetary gifts are: concert tickets, a bottle of wine, or a meal. Note that non-monetary gifts should only be recorded if they were used to compensate for the service. Non-monetary gifts that are given as personal tokens of appreciation should not be recorded.

OPEN-ENDED ANSWER
Q2_I. What was the amount of voluntary tip you paid? When filling in cents please enter a value from 00-99.

$[TEXT BOX].[TEXT BOX] [PROG- RANGE IS $0.01-1,000,000]

-99	Refused
-100	Valid Skip

SINGLE PUNCH ANSWER
Q2_J. Have you made any other transactions at a *hotel/motel* in the last calendar day?

00 No [SKIP TO Q3_A]
01 Yes
-99 Refused [SKIP TO Q3_A]
-100 Valid Skip

Instruction Page
Please record your next transaction in the same way as before. [PROCEED to new record for Q2_B]
[PN: MAX ITERATION FOR EACH SERIES OF QUESTIONS IS 20: WILL NOT SKIP TO NEXT ITERATION IF WE ALREADY COLLECTED 20 RESPONSES. INSTEAD, THEY SHOULD GO TO NEXT SECTION IN RANDOMIZATION].

SINGLE PUNCH ANSWER
Q3_A. In the last calendar day, have you made any transactions for *personal grooming, beauty, or massage services*?

00 No [SKIP TO Q4_A]
01 Yes
-99 Refused [SKIP TO Q4_A]

Instruction Page
On the next page, we will ask you to record one *personal grooming, beauty, or massage service* transaction that you made in the last calendar day. You will have an opportunity to record a separate transaction of this type later. If you had more than one transaction of this type (even at the same establishment and/ or during the same visit), please record each separately. Do not record transactions for which you have already provided information.

[NEXT]

SINGLE PUNCH ANSWER
Q3_B. What type of service did you receive? Record each transaction separately.

01 Hair Stylist
02 Barber
03 Manicurist/ Pedicurist
04 Massage Therapist
05 Waxing/ Hair Removal
06 Facial/ Skin Care
07 Makeup Artist
-99 Refused
-100 Valid Skip

///SOFT PROMPT///
IF RESPONDENT REFUSES Q3_B, STAY ON THE SAME PAGE AND WRITE "MISSING ANSWER(S)". IF THEY REFUSE AGAIN THEY SHOULD BE ALLOWED TO CONTINUE

SINGLE PUNCH ANSWER
Q3_C. Did you pay for this particular service (excluding any automatic or voluntary tip)?

00 No [SKIP TO Q3_F]

01 Yes
-99 Refused [SKIP TO Q3_F]
-100 Valid Skip

MULTIPLE PUNCH ANSWER
Q3_D. What payment type(s) did you use to pay your portion of the bill?
(select all that apply)

01 Cash
02 Debit
03 Credit
04 Check
05 Gift Card
06 Smartphone credit or app
07 Paper or online coupon (e.g., Groupon)
08 Other
-99 Refused
-100 Valid Skip

OPEN-ENDED ANSWER
Q3_E. What was the amount of the bill that you paid? When filling in cents please enter a value from 00-99.
(after tax, before automatic or voluntary tip)

$[TEXT BOX].[TEXT BOX] [PROG- RANGE IS $0.01-1,000,000]
-99 Refused
-100 Valid Skip

SINGLE PUNCH ANSWER
Q3_F. Did the business add an automatic tip for this service? If so, how much did you pay? When filling in cents please enter a value from 00-99.

00 No
01 Yes, and the amount was: $[TEXT BOX].[TEXT BOX]
-99 Refused
-100 Valid Skip

SINGLE PUNCH ANSWER
Q3_G. Did you leave a voluntary tip for this transaction?

00 No [SKIP TO Q3_J]
01 Yes
-99 Refused [SKIP TO Q3_J]
-100 Valid Skip

MULTIPLE PUNCH ANSWER
Q3_H. What payment type(s) did you use to pay the voluntary tip?
(select all that apply)

01 Cash
02 Debit
03 Credit
04 Check
05 Gift Card
06 Smartphone credit or app
07 Paper or online coupon (e.g., Groupon)
08 Non-monetary*
09 Other
-99 Refused
-100 Valid Skip

[Instructions at the bottom of response option list] *Examples of non-monetary gifts are: concert tickets, a bottle of wine, or a meal. Note that non-monetary gifts should only be recorded if they were used to compensate for the service. Non-monetary gifts that are given as personal tokens of appreciation should not be recorded.

OPEN-ENDED ANSWER
Q3_I. What was the amount of voluntary tip you paid? When filling in cents please enter a value from 00-99.

$[TEXT BOX].[TEXT BOX] [PROG- RANGE IS $0.01-1,000,000]
-99 Refused
-100 Valid Skip

SINGLE PUNCH ANSWER
Q3_J. Have you made any other transactions for *personal grooming, beauty, or massage services* in the last calendar day?

00 No [SKIP TO Q4_A]
01 Yes [PROCEED TO NEXT INSTRUCTION PAGE]
-99 Refused [SKIP TO Q4_A]
-100 Valid Skip

Instruction Page
Please record your next transaction in the same way as before. [PROCEED to new record for Q3_B] [PN:MAX ITERATION FOR EACH SERIES OF QUESTIONS IS 20: WILL NOT SKIP TO NEXT ITERATION IF WE ALREADY COLLECTED 20 RESPONSES. INSTEAD, THEY SHOULD GO TO NEXT SECTION IN RANDOMIZATION].

SINGLE PUNCH ANSWER
Q4_A. In the last calendar day, have you made any transactions for *moving or household maintenance services*?

00 No [SKIP TO Q5_A]
01 Yes

-99 Refused [SKIP TO Q5_A]

Instruction Page
On the next page, we will ask you to record one *moving or household maintenance service* transaction that you made in the last calendar day. You will have an opportunity to record a separate transaction of this type later. If you had more than one transaction of this type (even at the same establishment and/or during the same visit), please record each separately. Do not record transactions for which you have already provided information.
[NEXT]

SINGLE PUNCH ANSWER
Q4_B. What type of service did you receive? Record each transaction separately.

01 Professional Movers
02 Maid or Cleaning Service
03 Lawn/Gardening Service
04 Handyman/Repairman
05 Equipment Rental
-99 Refused
-100 Valid Skip

///SOFT PROMPT///
IF RESPONDENT REFUSES Q4_B, STAY ON THE SAME PAGE AND WRITE "MISSING ANSWER(S)". IF THEY REFUSE AGAIN THEY SHOULD BE ALLOWED TO CONTINUE

SINGLE PUNCH ANSWER
Q4_C. Did you pay for this particular service (excluding any automatic or voluntary tip)?

00 No [SKIP TO Q4_F]
01 Yes
-99 Refused [SKIP TO Q4_F]
-100 Valid Skip

MULTIPLE PUNCH ANSWER
Q4_D. What payment type(s) did you use to pay your portion of the bill?
(select all that apply)

01 Cash
02 Debit
03 Credit
04 Check
05 Gift Card
06 Smartphone credit or app
07 Paper or online coupon (e.g., Groupon)
08 Other
-99 Refused
-100 Valid Skip

OPEN-ENDED ANSWER

Q4_E. What was the amount of the bill that you paid? When filling in cents please enter a value from 00-99.
(after tax, before automatic or voluntary tip)

$[TEXT BOX].[TEXT BOX] [PROG- RANGE IS $0.01-1,000,000]
-99 Refused
-100 Valid Skip

SINGLE PUNCH ANSWER
Q4_F. Did the business add an automatic tip for this service? If so, how much did you pay? When filling in cents please enter a value from 00-99.

00 No
01 Yes, and the amount was: $[TEXT BOX].[TEXT BOX]
-99 Refused
-100 Valid Skip

SINGLE PUNCH ANSWER
Q4_G. Did you leave a voluntary tip for this transaction?

00 No [SKIP TO Q4_J]
01 Yes
-99 Refused [SKIP TO Q4_J]
-100 Valid Skip

MULTIPLE PUNCH ANSWER
Q4_H. What payment type(s) did you use to pay the voluntary tip?
(select all that apply)

01 Cash
02 Debit
03 Credit
04 Check
05 Gift Card
06 Smartphone credit or app
07 Paper or online coupon (e.g., Groupon)
08 Non-monetary*
09 Other
-99 Refused
-100 Valid Skip

[Instructions at the bottom of response option list] *Examples of non-monetary gifts are: concert tickets, a bottle of wine, or a meal. Note that non-monetary gifts should only be recorded if they were used to compensate for the service. Non-monetary gifts that are given as personal tokens of appreciation should not be recorded.

OPEN-ENDED ANSWER

Q4_I. What was the amount of voluntary tip you paid? When filling in cents please enter a value from 00-99.

$[TEXT BOX].[TEXT BOX] [PROG- RANGE IS $0.01-1,000,000]
-99 Refused
-100 Valid Skip

SINGLE PUNCH ANSWER
Q4_J. Have you made any other transactions for *moving or household maintenance services* in the last calendar day?

00 No [SKIP TO Q5_A]
01 Yes [PROCEED TO NEXT INSTRUCTION PAGE]
-99 Refused [SKIP TO Q5_A]
-100 Valid Skip

Instruction Page
Please record your next transaction in the same way as before. [PROCEED to new record for Q4_B] [PN: MAX ITERATION FOR EACH SERIES OF QUESTIONS IS 20: WILL NOT SKIP TO NEXT ITERATION IF WE ALREADY COLLECTED 20 RESPONSES. INSTEAD, THEY SHOULD GO TO NEXT SECTION IN RANDOMIZATION].

SINGLE PUNCH ANSWER
Q5_A. In the last day, have you made any transactions at a *casino*?

00 No [SKIP TO Q6_A]
01 Yes
-99 Refused [SKIP TO Q6_A]

Instruction Page
On the next page, we will ask you to record one *casino* transaction that you made in the last calendar day. You will have an opportunity to record a separate transaction of this type later. If you had more than one transaction of this type (even at the same establishment and/or during the same visit), please record each separately. Do not record transactions for which you have already provided information.

[NEXT]

SINGLE PUNCH ANSWER
Q5_B. What type of service did you receive? Record each transaction separately.

01 Dealers [SKIP TO Q5_F]
02 Floor Servers
03 Bar
04 Full-Service Dining (e.g., traditional restaurant)
05 Self-Service/Cafeteria/Buffets
06 Shuttle Service to/from Casino
07 Valet

-99 Refused [Continue to Q5_C]
-100 Valid Skip

///SOFT PROMPT///
IF RESPONDENT REFUSES Q5_B, STAY ON THE SAME PAGE AND WRITE "MISSING ANSWER(S)". IF THEY REFUSE AGAIN THEY SHOULD BE ALLOWED TO CONTINUE

SINGLE PUNCH ANSWER
Q5_C. Did you pay for this particular service (excluding any automatic or voluntary tip)?

00 No [SKIP TO Q5_F]
01 Yes
-99 Refused [SKIP TO Q5_F]
-100 Valid Skip

MULTIPLE PUNCH ANSWER
Q5_D. What payment type(s) did you use to pay your portion of the bill?
(select all that apply)

01 Cash
02 Debit
03 Credit
04 Check
05 Gift Card
06 Smartphone credit or app
07 Paper or online coupon (e.g., Groupon)
08 Other
-99 Refused
-100 Valid Skip

OPEN-ENDED ANSWER
Q5_E. What was the amount of the bill that you paid? When filling in cents please enter a value from 00-99.
(after tax, before automatic or voluntary tip)

$[TEXT BOX].[TEXT BOX] [PROG- RANGE IS $0.01-1,000,000]
-99 Refused
-100 Valid Skip

SINGLE PUNCH ANSWER
Q5_F. Did the business add an automatic tip for this service? If so, how much did you pay? When filling in cents please enter a value from 00-99.
00 No
01 Yes, and the amount was: $[TEXT BOX].[TEXT BOX]
-99 Refused

-100 Valid Skip

SINGLE PUNCH ANSWER
Q5_G. Did you leave a voluntary tip for this transaction?

00 No [SKIP TO Q5_J]
01 Yes
-99 Refused [SKIP TO Q5_J]
-100 Valid Skip

MULTIPLE PUNCH ANSWER
Q5_H. What payment type(s) did you use to pay the voluntary tip?
(select all that apply)

01 Cash
02 Debit
03 Credit
04 Check
05 Gift Card
06 Smartphone credit or app
07 Paper or online coupon (e.g., Groupon)
08 Non-monetary*
09 Other
-99 Refused
-100 Valid Skip

[Instructions at the bottom of response option list] *Examples of non-monetary gifts are: concert tickets, a bottle of wine, or a meal. Note that non-monetary gifts should only be recorded if they were used to compensate for the service. Non-monetary gifts that are given as personal tokens of appreciation should not be recorded.

OPEN-ENDED ANSWER
Q5_I. What was the amount of voluntary tip you paid? When filling in cents please enter a value from 00-99.

$[TEXT BOX].[TEXT BOX] [PROG- RANGE IS $0.01-1,000,000]
-99 Refused
-100 Valid Skip

SINGLE PUNCH ANSWER
Q5_J. Have you made any other transactions at a *casino* in the last calendar day?

00 No [SKIP TO Q6_A]
01 Yes [PROCEED TO NEXT INSTRUCTION PAGE]
-99 Refused [SKIP TO Q6_A]

-100 Valid Skip

Instruction Page
Please record your next transaction in the same way as before. [PROCEED to new record for Q5_B]
[PN: MAX ITERATION FOR EACH SERIES OF QUESTIONS IS 20: WILL NOT SKIP TO NEXT ITERATION IF WE ALREADY COLLECTED 20 RESPONSES. INSTEAD, THEY SHOULD GO TO NEXT SECTION IN RANDOMIZATION].

SINGLE PUNCH ANSWER
Q6_A. In the last calendar day, have you made any transactions for a *taxi, limousine, rideshare, or shuttle service*?

00 No [SKIP TO USRETH3]
01 Yes
-99 Refused [SKIP TO USRETH3]

Instruction Page
On the next page, we will ask you to record one transaction you have made for *taxi, limousine, rideshare, or shuttle service*. Do not record transactions for which you have already provided information. [NEXT]

SINGLE PUNCH ANSWER
Q6_B. What type of service did you receive? Record each transaction separately.

01 Limousine
02 Standard Taxi (e.g., "yellow cabs")
03 Uber, Lyft, or other Ride-Share service
04 Shuttle Service (e.g., Super Shuttle)
05 Valet
-99 Refused
-100 Valid Skip

///SOFT PROMPT///
IF RESPONDENT REFUSES Q6_B, STAY ON THE SAME PAGE AND WRITE "MISSING ANSWER(S)". IF THEY REFUSE AGAIN THEY SHOULD BE ALLOWED TO CONTINUE

SINGLE PUNCH ANSWER
Q6_C. Did you pay for this particular service (excluding any automatic or voluntary tip)?

00 No [SKIP TO Q6_F]
01 Yes
-99 Refused [SKIP TO Q6_F]
-100 Valid Skip

MULTIPLE PUNCH ANSWER
Q6_D. What payment type(s) did you use to pay your portion of the bill?
(select all that apply)

01 Cash

02 Debit
03 Credit
04 Check
05 Gift Card
06 Smartphone credit or app
07 Paper or online coupon (e.g., Groupon)
08 Other
-99 Refused
-100 Valid Skip

OPEN-ENDED ANSWER
Q6_E. What was the amount of the bill that you paid? When filling in cents please enter a value from 00-99.
(after tax, before automatic or voluntary tip)

$[TEXT BOX].[TEXT BOX] [PROG- RANGE IS $0.01-1,000,000]
-99 Refused
-100 Valid Skip

SINGLE PUNCH ANSWER
Q6_F. Did the business add an automatic tip for this service? If so, how much did you pay? When filling in cents please enter a value from 00-99.

00 No
01 Yes, and the amount was: $[TEXT BOX].[TEXT BOX]
-99 Refused
-100 Valid Skip

SINGLE PUNCH ANSWER
Q6_G. Did you leave a voluntary tip for this transaction?

00 No [SKIP TO Q6_J]
01 Yes
-99 Refused [SKIP TO Q6_J]
-100 Valid Skip

MULTIPLE PUNCH ANSWER
Q6_H. What payment type(s) did you use to pay the voluntary tip?
(select all that apply)

01 Cash
02 Debit
03 Credit
04 Check
05 Gift Card
06 Smartphone credit or app
07 Paper or online coupon (e.g., Groupon)
08 Non-monetary*

09 Other
-99 Refused
-100 Valid Skip

[Instructions at the bottom of response option list] *Examples of non-monetary gifts are: concert tickets, a bottle of wine, or a meal. Note that non-monetary gifts should only be recorded if they were used to compensate for the service. Non-monetary gifts that are given as personal tokens of appreciation should not be recorded.

OPEN-ENDED ANSWER
Q6_I. What was the amount of voluntary tip you paid? When filling in cents please enter a value from 00-99.

$[TEXT BOX].[TEXT BOX] [PROG- RANGE IS $0.01-1,000,000]
-99 Refused
-100 Valid Skip

SINGLE PUNCH ANSWER
Q6_J. Have you made any other transactions for a *taxi, limousine, rideshare, or shuttle service* in the last calendar day?

00 No [SKIP TO USRETH3]
01 Yes [PROCEED TO NEXT INSTRUCTION PAGE]
-99 Refused [SKIP TO USRETH3]
-100 Valid Skip

Instruction Page
Please record your next transaction in the same way as before. [PROCEED to new record for Q6_B]
[PN: MAX ITERATION FOR EACH SERIES OF QUESTIONS IS 20: WILL NOT SKIP TO NEXT ITERATION IF WE ALREADY COLLECTED 20 RESPONSES. INSTEAD, THEY SHOULD GO TO NEXT SECTION IN RANDOMIZATION].

USRETH3. Are you of Hispanic, Latino or Spanish origin?

- ☐ _1 Yes
- ☐ _2 No
- ☐ _3 Prefer not to answer

[Cortex 5 Standard Screener: DO NOT MODIFY]

[PN: USRETH3 must be asked before USRACE4]

USRACE4. What is your race?

Select all that apply.

- _1 White
- _2 Black or African American
- _3 Native American or Alaskan Native

- ☐ _4 Asian
- ☐ _5 Pacific Islander
- ☐ _6 Other race
- ☐ _7 Prefer not to answer [EXCLUSIVE]

[Cortex 5 Standard Screener: DO NOT MODIFY]

US01ETH (hidden question). Which of the following best describes you?

- ☐ White or Caucasian (not Hispanic or Latino)
- ☐ Black or African-American (not Hispanic or Latino)
- ☐ Asian/ Pacific Islander
- ☐ Native American, Alaska Native, Aleutian
- ☐ Hispanic or Latino (White or Caucasian)
- ☐ Hispanic or Latino (Black or African-American)
- ☐ Hispanic or Latino (all other races/ multiple races)
- ☐ Other
- ☐ Prefer not to answer

[Cortex 5 Standard Screener: DO NOT MODIFY]

US01ETH (mapping). US01ETH Mapping from USRETH3 & USRACE4

Hispanics are treated as 1st priority

- ☐ If USRETH3=1 (Yes)
- ☐ a) And if USRACE4= at least two punches (no matter which) OR USRACE4= only one punch among (3,4,5,6,7), then US01ETH= Hispanic or Latino (all other races/ multiple races)
- ☐ b) And if USRACE4= only one punch and that is 1, then US01ETH= Hispanic or Latino (White or Caucasian)
- ☐ c) And if USRACE4= only one punch and that is 2, then US01ETH= Hispanic or Latino (Black or African-American)
- ☐ Black are treated as 2nd priority
- ☐ If USRETH3= 2(No) or 3 (Prefer not to answer) and USRACE4=2 (no matter how many punches on race, as long one of them is =2), then US01ETH= Black or African-American (not Hispanic or Latino)
- ☐ Native American are treated as 3rd priority
- ☐ If USRETH3= 2(No) or 3 (Prefer not to answer) and USRACE4=3 (no matter how many punches on race, as long one of them is =3), then US01ETH= Native American, Alaska Native, Aleutian
- ☐ Asian are treated as 4th priority
- ☐ If USRETH3= 2(No) or 3 (Prefer not to answer) and USRACE4= 4 or 5 (no matter how many punches on race, as long one of them is =4 or 5), then US01ETH= Asian/ Pacific Islander
- ☐ Other are treated as 5th priority
- ☐ If USRETH3= 2(No) or 3 (Prefer not to answer) and USRACE4=6 (no matter how many punches on race, as long one of them is =6), then US01ETH= Other
- ☐ White are treated as 6th priority
- ☐ If USRETH3= 2(No) or 3 (Prefer not to answer) and USRACE4=1, then US01ETH = White or Caucasian (not Hispanic or Latino)
- ☐ Prefer not to answer are treated as 7th priority
- ☐ If USRETH3= 2(No) or 3 (Prefer not to answer) and USRACE4=7, then US01ETH = Prefer not to answer

The Paperwork Reduction Act requires that the IRS display an OMB control number on all public information requests. The OMB Control Number for this survey is 1545-2261. We estimate the time required to be eight minutes. Also, if you have any comments regarding the time estimates associated with this study or suggestions on making this process simpler, please write to:

Internal Revenue Service
Tax Product Coordinating Committee
1111 Constitution Avenue NW
Washington, DC 20224

Comparison of Estimates of Tipping Behavior Produced Using Probability and Non-Probability Samples: Methodology and Results

Prepared for Internal Revenue Service

Prepared by Fors Marsh Group LLC

November 2015

Version 1.2

Table of Contents

Summary

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 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.

Although the results of neither test clearly supported one sample being more biased than the other, the overall findings and considerations for the later, year-long fielding of the survey 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.

Introduction

The IRS intends to conduct a year-long survey of consumer tipping behavior, from here on referred to as the "Full Fielding", over the course of the 2016 calendar year. The potential target population for the IRS tipping study includes all U.S. residents who use services that are commonly tipped. The number of individuals in this population is unknown, but likely includes a majority of the U.S. adult population. Example settings where tipping is typical include: full-service restaurants, taxis, barber shops, beauty salons, hotels, and casinos.

The private nature of most transactions involving tipping makes it extremely difficult to collect reliable data that can be used to estimate total tip income. This difficulty is further compounded by the motivation of some individuals to not report tips received as taxable income. For these reasons, the IRS has concluded that surveying consumers about their tipping experiences is the most reliable way to collect quantitative data on tip income.

Prior IRS research on consumer tipping behavior found tipping rates varied considerably by industry and by region. A 1982 study conducted by the University of Illinois for the IRS [1] found tipping rates to be roughly 14% of the total bill for restaurants, 12% for barber and beauty shops, 19% for bars, and 20% for taxis. On a regional basis, mean restaurant tipping rates ranged from a low of 12.5% in the West North Central to a high of 15% in the Northeast.

The observed variation in tipping rates implies larger sample sizes are required in order to produce accurate estimates of tipping rates. Other things being equal, a larger sample size means greater cost. This constraint may be met in two ways: (1) limiting the scope of the study to focus on fewer industries/regions or (2) finding a more cost-effective mode of data collection. Due to the previous study's finding on the variance of tipping rates by industry and region, the IRS believes it would be inappropriate to limit the scope in these manners.

With respect to lowering the cost of data collection, an increasingly common alternative is the use of non-probability Internet samples.[2] The benefits of non-probability based panels relative to probability-based panels include:

1) The costs of sampling from an opt-in Internet panel may be substantially lower than the costs associated with sampling from a telephone- or mail-based frame, or a panel.
2) There might be costs or non-response associated with pushing individuals sampled from the telephone or mail frame to the Internet survey instrument, reflected in increased costs of sampling from Internet panels recruited from such frames (e.g., probability based web panel).[3]

[1] Pearl, R. B., & Sudman, S. (1983, June). *A survey approach to estimating the tipping practices of consumers* (Final Report to the Internal Revenue Service under Contract TIR 81-52); Pearl, R. B. (1985, July). *Tipping practices of American households: 1984* (Final Report to the Internal Revenue Service under Contract 82-21).
[2] Ansolabehere, S., & Schaffner, B. F. (2014). Does survey mode still matter? Findings from a 2010 multi-mode comparison. *Political Analysis*, *22*(3), 285-303.
[3] Dillman, D. A. (2013). Achieving synergy across survey models: mail contact and web responses from address-based samples. *Pacific Chapter of the American Association for Public Opinion Research*, *12*, 2013.

The chief drawback of using a non-probability sample from an Internet opt-in panel is that such panels could produce a realized sample that is less representative of the target population than the phone or mail frames. However, given the high rates of non-response associated with sampling from phone or mail frames, it is not clear to what degree respondents from probability samples are more representative with respect to tipping behavior than respondents contacted through an opt-in Internet panel, particularly after post-stratifying on observed demographic characteristics. Although non-response can be mitigated through follow-up contacts,[4] this exacerbates the differences between the probability and non-probability sampling strategies with respect to the cost of obtaining a sample of a given size, and such follow-up contacts have been shown to be associated with reductions in data quality[5]. Consequently, given a fixed budget it is unclear whether the reductions in bias in the estimates of mean tipping and stiffing rates that result from using a probability sample is worth the increase in the variability in these estimates that results from a smaller sample size, especially for relatively infrequent tipping transactions.

Given the uncertainty in the tradeoff between variance and bias in estimated tipping rates between a probability and non-probability sample, this consumer tipping study has followed Office of Management and Budget (OMB) guidelines [6] by conducting a pilot to resolve this conflict. Specifically, pilot surveys were fielded to a probability-based sample derived from the GfK KnowledgePanel and a non-probability based sample taken from Ipsos' s i-Say online opt-in panel over the course of July 2015 and responses were compared to determine if the results generated by two different Internet-based data streams produce equivalent estimates. This allows the IRS to estimate the degree to which there is a difference in bias that results from the use of a non-probability sample versus a probability sample. One benefit of using these two panels is that they both make use of a web-based interface which should reduce respondent burden, increase item response rates, and improve response accuracy compared to mail- or phone-based surveys.

<u>Non-probability Based Sample</u>: The Ipsos i-Say panel is an extensive opt-in research panel consisting of approximately 800,000 volunteers from across the United States. Individuals are recruited to participate on the panel from a variety of online sources, including numerous opt-in e-mail lists, banner and text links, and referral programs. Eligible participants who complete the study receive points that can be used toward charities, gift cards, or cash. Panelists who complete a screening questionnaire but do not qualify for the study also receive a small point-based incentive. Additionally, participants are entered into a monthly prize drawing. The monetary value of incentives for participation in this study is less than $1. Panelists represent a variety of ages, education levels, races, and ethnicities reflecting the diversity of the U.S. adult population. Invited panelists receive an e-mail with information about the study, and those who were interested follow a link to the study website where they answered a set of screening questions.

[4] Dykema, J., Stevenson, J., Klein, L., Kim, Y., & Day, B. (2013). Effects of e-mailed versus mailed invitations and incentives on response rates, data quality, and costs in a web survey of university faculty. *Social Science Computer Review*, *31*(3), 359-370.
[5] Olson, K. (2013). Do non response follow ups improve or reduce data quality?: a review of the existing literature. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *176*(1), 129-145.
[6] See Office of Management and Budget (2006). *Questions and answers when designing surveys for information collections*. Page 16, Section 22: "An agency may also use a pilot study to examine potential methodological issues and decide upon a strategy for the main study."

Probability Based Sample: The GfK KnowledgePanel is an Internet panel that uses a probability-based sampling strategy where the survey frame is derived from the USPS Delivery Sequence File and is therefore representative of the US adult population. Individuals are invited to participate in the panel by mail, followed by telephone calls for those who do not respond to the initial invitation. For those individuals selected for participation without computers or an Internet connection, a netbook is provided. This process attempts to mitigate the selection bias associated with web surveys while preserving the benefits associated with a computer interface.

A benefit of the KnowledgePanel relative to the opt-in panel is that knowing the probability of selection allows researchers to estimate total survey error. The ability to estimate total survey error would in theory allow for the calculation of unbiased estimates of tipping behavior from a probability-based sample if non-response is random conditional on observable covariates. However, if estimates derived from the Ipsos and GfK samples support statistically indistinguishable conclusions about the tipping behavior across industries and geographic areas, we would recommend using the more cost-efficient non-probability based method. If identical, the use of the i-Say panel would generate more usable data at lower cost than would a probability-based sample, without a substantial decrement to the accuracy of the tipping estimates.

The next section describes the methodology used to compare the probability and non-probability panels with respect to the representativeness of respondent tipping behavior.

Methodology

The current section describes two methodologies that will be used to decide between probability and non-probability samples for the Full-Fielding of the consumer tipping survey. The first method involves testing for differences in tipping behavior between individuals sampled from probability and non-probability panels, assuming that the non-probability sample is at least as biased with respect to population tip rates as the probability sample and less costly per completed survey. The second methodology involves comparing tipping behavior of individuals sampled from both panels to estimated mean tip rates derived from Point of Sale (POS) data, assuming that the POS data is no more biased than either survey-based sample.

"Differences in Samples" in Tipping Behavior Between Probability and Non-Probability Panelists

As discussed in the introduction, the GfK KnowledgePanel represents a benchmark because of its combination of a representative frame and probability sampling from that frame. Under the assumption that an estimate derived from a probability sample is at least as accurate as that derived from a non-probability sample with respect to tipping behavior, then the choice of whether to use the probability or non-probability sample is reduced to the well-known bias versus variance trade-off in statistics. The bias vs. variance trade-off in statistics states that, given the same sample, decreases in bias/ increases in accuracy in an estimate come at the cost of increases in the uncertainty about that estimate. To add a little context, statistical interventions to increase accuracy oftentimes come at the expense of statistical certainty, as the intervention usually attempts to more closely conform to the data, which may not work quite the same in another sample —a notion that is built into the estimate. However, given that we are comparing different samples (i.e., not the same sample with different estimation interventions), and we know that the cost per completed survey will be lower with the non-probability sample, then if the samples do not differ with respect to tipping behavior (i.e., are equally accurate), the non-probability sample can be said to be superior because of the larger potential sample size, and thus lower degree of sampling-related error (i.e., lower variance/ uncertainty) in the final estimates. To test for similarities in tipping behavior between the two samples, what will subsequently be referred to as a "Difference in Samples" test, the Fors Marsh Group (FMG) team can estimate the following models:

1) [illegible]

In Equation 1, $\hat{T}_{tijs}$ is a tip rate greater than 0 of full-service restaurant transaction *t* for respondent *i* residing in location *j* and sample *s*; Ipsos is an indicator variable that takes a value of 1 if the respondent was part of the Ipsos i-Say panel and 0 if part of the GfK KnowledgePanel. Equation 1 allows for a test of an unconditional difference in tipping rates, i.e., systematic differences in tipping rates between the samples that can be driven by differences in either observed or unobserved demographic or geographic characteristics of respondents in the two samples. Specifically, a δ that is significantly different from 0 is consistent with unconditional differences in behavior between respondents from the two samples. Because of the small number of estimated parameters (k=2) of this model, it allows for precise / low-error estimates of this unconditional difference even with small samples. Additionally, the test for bias in the non-probability sample can be made robust to violation of the assumption of equal variances in both samples through the use of robust standard errors.

Another potential concern is that the differences are not independent across transactions or individuals due to the fact that multiple respondents may visit similar restaurants. To account for this, standard errors for each test are clustered at the level of the commuting zones, an aggregation of counties which send and receive large fractions of their resident working populations to each other but not to counties in other commuting zones.[7] Commuting zones have been used in recent, prominent studies to define the geographic extent of environmental determinants of social outcomes.[8] Commuting zones may proxy for the typical geographic extent of respondents' daily travels, and thus the restaurants they are likely to visit. To the degree that unobserved restaurant characteristics are systematically related to tip rates, and given that respondents in the same commuting zones may visit the same restaurants, tip rates for respondents in the same commuting zone may be more similar than tip rates for respondents in different commuting zones. Clustering the standard errors at the commuting zone level will account for any effect on sampling variability that results from localized, unobserved restaurant sector effects on the outcomes of interest.[9]

Given that we can use sample weights provided by both vendors to calibrate the results from the final fielding and our own frame to match the demographic and geographic characteristics of our population of interest, the IRS is interested in differences in tipping behavior between the two samples not explained by differences in observable demographic characteristics. Consequently, we may wish to estimate conditional differences in the tip rate between the two models, i.e., the differences in tipping behavior attributable to unobserved differences between the two samples. Specifically, we can estimate the following model separately:

2) □̂□ = □□□□□ + □□□ + □□□ + □□□□□□□□□□□□□□□□

In Equation 2, □ is a vector of demographic characteristics of person *I* observable in both samples as well as in the 5-year 2013 American Community Survey (ACS) that will likely be used to construct our frame to weight to the Full-Fielding; and □ is a vector of geographic characteristics of area *j*. See Table 1 in the Appendix for variable descriptions. If parameter δ is significantly different from zero and at least one parameter within □ or □ is also significantly different from 0, then the estimated model is consistent with a conditional difference in tipping rates between the two samples (if δ is significantly different from zero but □ and □ are not, this collapses to an unconditional difference in tipping rates between the two samples).

[7] Tolbert, C. & Sizer, M. (1996). U.S. Commuting Zones and Labor Market Areas: A 1990 Update. ERS Staff Paper Number 9614. Economic Research Service, Rural Economy Division, U.S. Department of Agriculture, Washington, D.C.
Note: We use commuting zone definitions for the year 2000, the last year for which the USDA has produced commuting zone definitions. Source: http://www.ers.usda.gov/data-products/commuting-zones-and-labor-market-areas/documentation.aspx

[8] Chetty, R., Hendren, N., Kline, P., & Saez, E. (2014). Where is the land of Opportunity? The Geography of Intergenerational Mobility in the United States. *The Quarterly Journal of Economics*, *129*(4), 1,553-1,623.

[9] Cameron, C. & Miller, D. (2015). A Practitioner's Guide to Cluster-Robust Inference. *Journal of Human Resources*, *50*(2), 317-373.

"Differences in Differences" in Tipping Behavior Between Probability and Non-Probability Panelists and POS data

Although the first part of the proposed analysis of the pilot survey data assumes that a sample from the GfK KnowledgePanel yields estimates that are as accurate as estimates derived from the Ipsos i-Say panel, the validity of using the probability estimates as a benchmark is compromised if this assumption does not hold. For example, it might be the case that individuals who join opt-in Internet panels (e.g., i-Say panelists) do not differ from the general population with respect to tipping, but those who to respond to solicitations through the mail (and thus participate in GfK's KnowledgePanel) do. In essence, there's a possibility of some unknown tipping difference between people who join panels using the mail and online. To examine whether the conclusions drawn from the first part of the analysis still hold when relaxing this assumption, probability and non-probability estimates of tipping rates are compared with estimates derived from POS data.

We assume of the POS data that the transactions represented are an accurate estimate of the "true" mean tipping rate. Because the restaurants represented in the data attempt to accurately record all tipping transactions, POS data is less likely to suffer from potential social desirability biases in reported tip rates (i.e., remembering tipping more on a transaction than one actually did). However, our accuracy assumption may be violated if there is systematic misreporting in tip amounts or bill sizes in the POS data or if establishment mean tipping rates are systematically related to the propensity of the restaurant to report POS data. The report *An Assessment of the Validity of Using Point-of-Sale Data to Estimate Restaurant Tipping Rates* [10] discusses the possibility of measurement error with respect to transactions for which the tips were paid with cash and the potential for measurement error in the bill size for transactions utilizing forms of prepayments (e.g., Groupon). Consequently, using the POS data as a benchmark will likely only be valid for non-cash, non-prepaid transactions. This represents a difference from the "Difference in Samples" test, which involved a comparison of the mean tip rate for transactions involving all forms of payment at full-service restaurants. The POS validation report also found issues with respect to establishment "non-response." Specifically, there were too few tipping transactions in establishments identified as quick-service establishments (i.e., those that did not provide table service to customers) to estimate a reliable tip rate for those establishments. Thus, POS data can only be used as a baseline for full-service restaurants. Although the report found little evidence of systematic differences in establishment representation across Designated Market Areas (DMAs), there was no ability to test for differential establishment inclusion within DMAs. These issues may undermine the reliability of the POS-derived estimates of mean tip rates in our population of interest. Consequently, the "Differences in Differences" analysis is not necessarily more informative or better than the "Differences in Sample" analysis, but rather complementary with its own strengths and weaknesses.

To estimate the unconditional "Differences in Differences," we estimate the following model:

3) [illegible]

[10] *An Assessment of the Validity of Using Point-of-Sale Data to Estimate Restaurant Tipping Rates* (2014). Internal report prepared for the Internal Revenue Service by Fors Marsh Group under contract TIRNO-13-Z-00021-0002.

3[illegible]) |[illegible] − [illegible]| = [illegible] + [illegible]

Similarly, to estimate the conditional "Differences in Differences," we estimate the following model :

4[illegible]) [illegible] − [illegible] = [illegible] + [illegible] + [illegible] + [illegible]

4[illegible]) |[illegible] − [illegible]| = [illegible] + [illegible] + [illegible] + [illegible]

The left-hand side of both Equations 3 and 4 are deviations of a survey transaction tip rate from the estimated average tip rate implied by the POS average ($\overline{T}_{jPOS}$) for the transaction's geographic unit (i.e., commuting zone). Controlling for the geographic average tipping rate for the POS transactions by subtracting it from the left-hand side allows for the incorporation of individual-level predictors.

Using Equations 4, however, changes the interpretation of δ. Under Equations 4, δ is the marginal effect of being in the Ipsos (versus GfK) sample on the deviation of the reported tip rate from the commuting zone average. Note that previously (i.e., in Equations 1 and 2) δ referred to the marginal effect of being in the Ipsos (versus GfK) sample on the tip rate. Equations 4a and 4b are then models of within-geographic-unit selection bias if we assume the POS data as the gold standard. Hence, to the extent that Ipsos or GfK differs less from the POS data, that sample appears to be more accurate and should be preferred. Specifically, we require first that δ be significantly different from 0. If δ is significantly different from 0, if the predicted absolute mean deviation of the Ipsos sample tip rate from the local POS average tip rate is larger than for the GfK tip rate, then the GfK sample tip rate will be preferred or vice versa.

We refer to Equations 3a and 4a as the "Differences in Differences" tests as they allow for a test of differences in the systematic deviation of respondents between samples in the same direction across geographic units. By contrast, we refer to 3b and 4b as "Differences in Absolute Differences" tests which allow the direction of the deviations to vary across commuting zones. We argue that Equations 3a and 4a may be more useful for determining relative bias of the panels for the national mean tipping rate; however, we argue that 3b and 4b may be more useful for testing for relative bias and/or sampling variance at the local level.

The difference in focus between the difference in difference and the difference in absolute difference is important if the IRS desires to develop small area estimates of tipping rates as Equations 3b and 4b reflects the differences in the degree of dispersion around the local area average tip rate between different samples and strata. Consequently, if for example, the Ipsos sample has a larger absolute deviation than the GfK sample, that may indicate that local area estimates of the tipping derived from the Ipsos sample will suffer to a greater degree from sampling variability and thus potentially unreliability and uncertainty, though it does not necessarily indicate systematic bias, as the mean tipping rate may be close to the true local area tipping rate if the local area sample is sufficiently large. This variability may in practice be mitigated by using model-assisted approaches to impute local area estimates of the mean tipping rates, such as multilevel regression and poststratification (MRP)[11], which utilize information from the entire sample, rather than just information from respondents in the local area, to estimat e the local-area's mean tipping rate, thus limiting the effect of sampling variability on the local area estimates. The "Differences in Absolute Differences" test may consequently be less relevant with respect to adjudicating between the samples if 1) the primary interest is in the national tipping rate or 2) model -assisted methodologies are used to generate local area estimates.

Given that $\overline{T}_{jPOS}$ is subject to sampling error (as it built from many transactions per commuting zone), we will cluster the estimated standard errors at the level of the commuting zones to account for the automatic correlation in residuals that the inclusion of $\overline{T}_{jPOS}$ on the left hand side induces across units in the same commuting zon e due to the use of the same/ similar businesses and other local area characteristics.

In summary, t he focal null hypothesis for the "Differences in Differences" tests then becomes:

$$5)\ |\,(\hat{\ } - \bar{\ }\,|\ = 1)| = |\,(\hat{\ } - \bar{\ }\,|\ = 0)|$$

Equation 5, when applied to equations 3 a/ b and 4a/ b, tests the extent to which the expected value/ mean difference from the POS data for the Ipsos sample is the same as the expected value/ mean difference from the POS data for the GfK sample —a null hypothesis significance test which can be evalua ted using the well -known Wa ld Test from a maximum likelihood estimate . Based on the assumptions discussed earlier, we would interpret the sample with the smaller absolute average distance from the POS mean as being less biased , more accurate, and the preferred vendor.

Rules for Deciding Between the Probability and Non -Probability Samples

Once the results of the "Differences in Samples" and "Differences in Differences" tests have been obtained, a methodology is required to aggregate all the results in such a way that an inference can be drawn concerning whether to sample from the probability or non -probability panels. Table 1 presents some potential decision rules. The outcome space represents a clear simplification insofar as multiple variants (tip rat e versus conditional versus unconditional tests; using weights) of these "Differences in Samples" and "Differences in Differences" tests are likely to be implemented for the purpose of evaluating how well the tests hold up to generally minor changes in app roach.

However, assuming that results are consistent for each set of tests, Table 1 reflects the following decision rule: if either test indicates that the probability sample is less biased than the non - probability sample, then the FMG Team will recommen d using the probability sample for the Full-Fielding; otherwise, the FMG Team will recommend the use of the non -probability sample. The rule is

[11] See *Buttice, M. K., & Highton, B. (2013). How Does Multilevel Regression and Poststratification Perform with Conventional National Surveys.? Political Analysis, 21(4), 449-467.* for a description of MRP and a test of its sampling properties.

a result of the continued skepticism of non -probability samples among many survey statisticians .[12] This rule is may be especially valid with respect to bias in estimates for establishments other full-service restaurants where the bill or tip was paid non -electronically. The second rule is based on the assumed lower cost of the non-probability sample, which, assuming comparable levels of estimate accuracy, will naturally determine the decision. Also note that this rule assumes that reducing response bias is more important than reducing variability.

Table 1 – Decision Matrix – Probability Sample as "Gold Standard"

		"Differences in Differences" Test Result		
		Probability	Neither Probability Nor Non-Probability	Non-Probability
"Differences in Samples" Test Result	Probability	*Probability*	*Probability*	*Probability*
	Neither	*Probability*	*Non-Probability*	*Non-Probability*

Note: Rows and columns reflect the sampling strategy with less bias based on the result of the test. Italicized options represent the sampling strategy that will be recommended depending on the given constellation of the two tests

Depending on one's belie fs, different decision rules are possible. For example, if one believed that (1) there is no theoretical basis to believe that the probability sample suffers from less selection bias than the non-probability sample, (2) the POS data was more reliable than survey data because of social desirability issues, and (3) that differences in bias in reported tip rates for full -service restaurants was likely to carry over to other industries, then we may instead prefer the following decision matrix:

Table 2 – Decision Matrix – Probability Sample Not "Gold Standard"

		"Differences in Differences" Test Result		
		Probability	Neither Probability Nor Non-Probability	Non-Probability
"Differences in Samples" Test Result	Probability	*Probability*	*Non-Probability*	*Non-Probability*
	Neither	*Probability*	*Non-Probability*	*Non-Probability*

Note: Rows and columns reflect the sampling strategy with less bias based on the result of the test. Italicized options represent the sampling strategy that will be recommended depending on the given constellation of the two tests.

[12] AAPOR (2013). "Report of the AAPOR Task Force on Non -Probability Sampling." https://www.aapor.org/AAPORKentico/AAPOR_Main/media/MainSiteFiles/NPS_TF_Report_Final_7_revised_FNL_6_22_13.pdf

Consequently, there may be no "objective" means to map the results of the "Differences in Samples" and "Differences in Differences" tests to a decision. It may still be useful to lay out one's assumptions and resulting decision rules before the actual empirical analysis is undertaken in order to avoid the biases that can result from post -hoc rationalization. In drawing inference from the results reported in the next session, we will utilize both matrixes in order to assess the robustness of our findings.

Data

The data collected for the purpose of the analysis from the two samples consists of bill sizes and tip amounts for 1,832 full service restaurant transactions undertaken by 12,137 respondents in the 24 hours before undertaking the survey. In addition, both surveys included information on respondent demographics (X_i) including, age, gender, educational attainment, race/ethnicity, and household income. Both vendors also provided the respondent's zip code, which allowed relevant, primarily county-level geographic information (G_j) to be appended, including the percentage of the respondent's county which was foreign born (5 -year ACS), the size of the metropolitan area in which the respondent resides, urban/rural status of the respondent's county (USDA), and census division.

Descriptive statistics for the raw samples for the GfK and Ipsos samples, respectively, are reported in Tables 9 and 11 in the Appendix. We begin by noting that these descriptive statistics reveal differences between the Ipsos and GfK samples on several characteristics. We formally test for imbalance in these characteristics in the raw samples in the first and third Columns of Table 15. Both the linear and logit models indicate that many demographic and geographic variables predict sample membership which suggests slightly different compositions in the Ipsos and GfK samples and the importance of controlling for such differences in the "Differences in Sample" and "Differences in Differences" tests.

It is important to note that for the "Differences in Samples" and "Differences in Differences" performed on this raw sample to be valid, we must assume that tipping behavior does not systematically differ across different groups defined by the demographic and geographic characteristics; such an assumption may not be realistic. For example, it might be the case that individuals with Internet access in rural areas are more likely to be overrepresented in the Ipsos frame relative to GfK and, in addition, differ to a greater extent with respect to tipping behavior from the average rural resident. By contrast, individuals with Internet access in urban areas many not differ from the average urban resident, due to the more widespread access to and use of the Internet in urban areas, and may be more evenly represented in both samples. The imbalance in rural residents is likely, however, to result in bias in the estimates.

This assumption of a constant difference in mean tipping rates between the two samples observed in the results of Table 3 is based solely on the obtained sample and is not necessarily problematic if the weighted estimation samples are representative of the target population with respect to these relevant background characteristics. Bias is avoided if each sample is derived from the same population because the estimate of δ (i.e., the between sample difference) will still represent the average difference in the population. However, if the pooled unweighted estimation sample differs from one another with respect to characteristics relevant to the tip rate, then our evidence suggests that δ will not be sample differences from the same population, but rather represents of the difference in the population estimate one would obtain from the two samples, and would thus be biased.

We address the potential for bias in the estimates derived from the raw samples by re-estimating all "Differences in Samples" and "Differences in Differences" using sample weights. The sample weights we used were post-stratification weights provided by both the Ipsos and GfK vendors. We would like to find evidence that both vendors have designed their survey weights to ensure that, when weighted, samples are representative of the same, appropriate target population (all adults residing in the United States). Importantly, we would like to find evidence suggesting that, when considering relevant sample characteristics, the weighted samples do not look substantially different. If the samples do not appear to be different on important characteristics, then the estimate of δ obtained from the pooled, weighted sample should not be biased substantially.

Evidence suggesting that both weighted samples represent a similar population can be observed in Table 15. Specifically, Table 15 shows the differences between the unweighted and weighted regression models which predict sample membership using observable demographic and geographic variables. Columns 1 and 3 represent the unweighted samples, which show several differences across samples. In particular, there is an increase in the probability of being part of the Ipsos sample (versus GfK) when younger, less educated, an ethnic minority, and making less income. When comparing the results in column 2 and 4 (representing the weighted samples) to the unweighted results, the coefficients for age, education, race/ethnicity, and income categories are all substantially reduced (but not eliminated). Moreover, the model fit comparing weighted to unweighted samples changes substantially (dropping by about half). Taken together, we argue that the pattern is consistent with the vendor weights making both samples more representative of the same population, though there is still some degree of imbalance. The potential bias in δ should be kept in mind when interpreting the results.

One limitation worth noting when incorporating the sample weights is that sample weights often result in an increase in sampling variability/standard errors for reductions in bias, resulting in reduced statistical power. Consequently, for the purpose of robustness, results are reported for each test using both the weighted and unweighted sample.

Results

In the coming section we present results for the "Differences in Samples" and "Differences in Differences" tests for the set of full-service restaurant[13] transactions with a fully voluntary gratuity[14] obtained from the GfK and Ipsos samples.

Table 3 – Estimates of Average Differences in Ipsos and GfK (□) by Test

	Unconditional Differences in Sample	Conditional Differences in Sample	Unconditional Differences in Differences	Conditional Differences in Differences	Unconditional Differences in Absolute Differences	Conditional Differences in Absolute Differences
□□	-0.004 (-0.003)	-0.006 (0.003)*	-0.003 (-0.003)	-0.005 (-0.003)	0.003 (-0.002)	0.006 (0.002)*
Control Variables?	No	Yes	No	Yes	No	Yes

Robust standard errors clustered on Commuting Zones in parentheses. * $p<0.05$; ** $p<0.01$

Table 4 – Estimates of Average Differences in Ipsos and GfK (□) by Test, Weighted

	Unconditional Differences in Sample	Conditional Differences in Sample	Unconditional Differences in Differences	Conditional Differences in Differences	Unconditional Differences in Absolute Differences	Conditional Differences in Absolute Differences
□□	-0.002 (-0.003)	-0.004 (-0.003)	-0.001 (-0.004)	-0.003 (-0.004)	0.003 (-0.002)	0.005 (0.002)*
Control Variables?	No	Yes	No	Yes	No	Yes

Robust standard errors clustered on Commuting Zones in parentheses. * $p<0.05$; ** $p<0.01$

"Differences in Samples" Test

The initial, unconditional "Differences in Samples" (Equation 1) test results are reported in the first columns of Table 3 and 4. The estimated mean Ipsos tipping rate is approximately 0.4 percentage points lower than the GfK tipping rate in the unweighted sample and 0.2 percentage points lower in the weighted sample. This difference is not statistically significantly different from zero. Hence, under the assumption that the GfK estimate represents a "gold standard," the result of the unconditional "Differences in Samples" test is consistent with the Ipsos estimate being unbiased, and thus favors the use of the Ipsos sample.

We also estimated the conditional model (Equation 2) in column 2 of Tables 3 and 4 which adds the individual-level and geographic control variables to account for observable differences between the respondents in the two samples[15]. The point estimate for the conditional difference is 0.6

[13] This definition includes both free-standing restaurants as well as those housed in a casino or hotel.
[14] Due to the high degree of measurement error apparent in responses to the automatic gratuity amount, all observations with an automatic gratuity were excluded from the analysis.
[15] Some observations are lost from the Ipsos sample in column 2 due to missing values for the control variables. To examine the degree to which these dropped observations may affect the inference regarding the difference in tipping between Ipsos and GfK, in Table 5 the unconditional tests are run for the subsample with no missing observations on the

percentage points and statistically significantly different from zero at the 5% level, with GfK respondents reporting higher tipping rates conditional on the observables. Thus, the differences in composition of the samples appeared to mask possible differences between GfK and Ipsos on their average tipping rate. The result from the conditional "Differences in Samples" test favors the use of the GfK sample. However, in the conditional differences in sample test for the weighted sample, the difference between Ipsos and GfK is now not statistically significant. As previously noted, the use of sample weights may result in an increase in sampling variability/standard errors for reductions in bias, resulting in reduced statistical power. However, the loss of significance in the conditional differences in sample test appears to be due to the reduction in the size of the coefficient (from approximately 0.6 percentage points to 0.4 percentage points) rather than an increase in variability, as indicated the stability in the size of the standard error.

"Differences in Differences" Test

We then moved on to the "Differences in Differences" test, where the dependent variable is the difference between the tipping rate for a transaction and the mean commuting zone tipping rate computed using the point of sale data. The results of the unconditional "Differences in Differences" test (Equation 3a) are reported in the third column of Tables 3 and 4. The unconditional difference in difference is not statistically significant and shows a 0.3 percentage point estimated difference between Ipsos and GfK samples for the unweighted sample and a 0.1 percentage point difference in the weighted sample. The unconditional "Differences in Differences" test, like its "Differences in Samples" counterpart, thus supports the use of the Ipsos sample.

We next estimated a conditional "Difference in Difference" model (Equation 4a) including control variables. As compared to the "Differences in Samples" test, the conditional "Differences in Differences" test is not statistically significant as is depicted in column 4 of Tables 3 and 4 with a 0.5 percentage point difference between Ipsos and GfK in the unweighted sample and a 0.3 percentage point difference in the weighted sample.

In addition to the "Differences in Differences" tests, we also evaluated difference in the absolute difference between the tip rate and the commuting zone averaged tip rate (i.e. Equation 3/4b) in column 5 (unconditional) and 6 (conditional). The differences in absolute differences mirrored the results from the "Differences in Samples" tests as the unconditional differences in absolute differences was not significantly different from zero, yet was statistically significantly different for the conditional differences in absolute differences test obtaining a 0.6 percentage point difference between Ipsos and GfK in the unweighted sample and a 0.5 percentage point difference in the weighted sample. To the degree that this difference in the absolute difference indicates that there would be greater bias/variability in local area estimates derived from the Ipsos sample, this result would argue in favor of using GfK.

Interestingly, a reduction in the size of the Ipsos coefficient is observed across all tests, consistent with the differences in the sample mean tip rates between being larger than the differences one

[15] (cont.) control variables. The estimated unconditional difference as well as the standard errors are very similar to the full estimation sample, consistent with little systematic difference between missing and complete cases with regards to tipping.

would find if the sample were representative of the general population. In the Full-Fielding, an additional post-stratification effort will be undertaken to ensure that the sample matches the population with respect to tipping-relevant demographic and geographic characteristics.

Implications of the Results for Deciding Between the Probability and Non-Probability Samples

Given the results of all weighted and unweighted tests, we can proceed to making a recommendation as to the panel to choose for the final fielding. We make the recommendation by using the decision matrices outlined in the previous section. The evidence from the "Difference in Samples" tests is as follows:

a) All unconditional "Differences in Sample" tests found little evidence of systematic differences in the tipping rates between the GfK and Ipsos samples.
b) The conditional "Differences in Sample" was statistically significant when using an unweighted sample.
    a. The significant result was not robust to weighting the combined sample such that it is more representative of the target population.
    b. The size of the difference between the sample tip rates was also generally small (0.2 to 0.4 percentage points).
    c. Assuming GfK represents a "gold standard," our findings show little to no bias in the estimates of the mean tip rate obtained from the Ipsos data.

The "Differences in Sample" tests consequently provides support for *neither the Ipsos nor GfK* sample when it comes to final fielding.

The evidence from the "Difference in Differences" tests is as follows:

c) All "Differences in Differences" test results showed no systematic difference in the tipping rates between the GfK and Ipsos samples.
d) All unconditional differences in absolute differences tests showed no systematic differences in tipping rates between the GfK and Ipsos samples.
e) All conditional differences in absolute differences tests showed systematic differences in tipping rates between the GfK and Ipsos samples.
    a. The absolute difference between a respondent's reported tip rate and the commuting zone average is higher for Ipsos respondents when incorporating controls.
    b. As discussed in the Methodology section, the conditional difference in absolute difference result is not unequivocal evidence that the national or local estimates for the mean tipping rate will be more biased for the Ipsos sample than for the GfK.

We interpret the above evidence to show that the "Differences in Differences" test supports *neither the probability nor non-probability samples*.

Table 5 – Decision Matrix – Probability Sample as "Gold Standard"

		"Differences in Differences" Test Result		
		Probability	Neither Probability Nor Non-Probability	Non-Probability
"Differences in Samples" Test Result	Probability	*Probability*	*Probability*	*Probability*
	Neither	*Probability*	*Non-Probability*	*Non-Probability*

Table 6 – Decision Matrix – Probability Sample Not "Gold Standard"

		"Differences in Differences" Test Result		
		Probability	Neither Probability Nor Non-Probability	Non-Probability
"Differences in Samples" Test Result	Probability	*Probability*	*Non-Probability*	*Non-Probability*
	Neither	*Probability*	*Non-Probability*	*Non-Probability*

To summarize, given the evidence outline above, both decision matrices above would support the use the Ipsos sample, given the lower cost per completed survey, and thus a larger sample and the resulting potentially more precise estimates of the tip and stiffing that can be obtained from that vendor, especially for non-full service restaurant industries.

Summary and Conclusions

The current report describes methodologies that can be used to decide between the use of probability and non-probability panels for the purpose of generating a sample of respondents for the consumer tipping survey. Specifically, the methodologies outlined allow for a test of differences in selection and/or response bias between these panels. The first method, termed the "Differences in Samples" test, assumes that the probability sample is no more biased than the non-probability sample. Consequently, any difference in reported (conditional or unconditional) average tip rates between the two samples is interpreted as indicating bias in the non-probability sample. By contrast, the "Differences in Differences" test does not make this assumption and utilizes information about tipping transactions from POS data as an objective arbiter between the probability and non-probability samples.

Although the results of neither test clearly support one sample being more biased than the other, we recommend the use of the Ipsos 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 Ipsos sample is preferable. Therefore, the Fors Marsh Team recommends that the IRS field the final survey to the Ipsos non-probability panel.

Appendix

Data Cleaning

We observed several instances of extremely high bill amounts, tip amounts, and tip rates in the survey data. Assuming some the unusual and unexpected data points represent measurement error or invalid transactions, a n outlier identification strategy simil ar to that employed in the report *An Assessment of the Validity of Using Point -of-Sale Data to Estimate Restaurant Tipping Rates* can be employed.

Specifically, we assume that bill size and tip amount are log normally distributed and tip rate is normally distributed for each transaction type (e.g. , full service restaurants, hair dressers) [16]. For both the Ipsos and GfK sample, we then calculate the following ratio for each outcome by transaction type as follows:

$$\frac{|y - Q_{75}|}{Q_{75} - Q_{25}} \text{ for } y > Q_{75}$$

$$\frac{|y - Q_{25}|}{Q_{75} - Q_{25}} \text{ for } y < Q_{25}$$

Where y is logged bill amount, logged tip amount, or tip rates. Transactions are identified as outliers if either ratio exceeds 2.5 for bill amount, tip amount, or tip rates. Respondents with at least one outlier transaction are excluded from the analysis. Descriptive statistics for t he full service restaurant transactions reported by these excluded individuals are reported separately for GfK and Ipsos respondents in Table s 7 and 8.

Descriptive Statistics

Table 7 – Descriptive Statistics for Outlying Full Service Restaurant Transaction s - GfK Sample Excluded Outliers

Variable	*N*	Mean	Standard Deviation	Minimum	Maximum
Bill Amount	68	$268.48	$858.16	$1.00	$5639.00
Tip Amount	72	$86.07	$223.43	$0.00	$1100.00
Was Transaction Tipped?	64	0.97	0.18	0.00	1.00
Tip Rate	57	146.00%	456.81%	0.15%	2500.00%

[16] We recognize the normality assumption applied may not hold due to non -independence of transactions within commuting zones as well as individual respondents. However, the small number of transactions per commuting zone and individual makes identifying outliers by commuting zone and individual unfeasible.

Table 8 – Descriptive Statistics for Full Service Restaurant Transactions - Ipsos Sample Excluded Outliers

Variable	*N*	Mean	Standard Deviation	Minimum	Maximum
Bill Amount	194	$959.54	$7111.45	$0.44	$75000.00
Tip Amount	189	$849.56	$7190.32	$0.00	$75000.00
Was Transaction Tipped?	96	0.83	0.37	0.00	1.00
Tip Rate	74	90.82%	191.54%	0.12%	1608.62%

Table 9 – Unweighted Descriptive Statistics - GfK Sample

Respondent-Level Variables	*N*	Mean	Standard Deviation	Minimum	Maximum
Full Service Restaurant Transactions in Last Day	5,663	0.20	0.44	0.00	4.00
Male	5,663	0.49	0.50	0.00	1.00
Age, Excluded Category = *18-24*					
25-34	5,663	0.16	0.37	0.00	1.00
35-44	5,663	0.15	0.35	0.00	1.00
45-64	5,663	0.39	0.49	0.00	1.00
65+	5,663	0.22	0.42	0.00	1.00
Age, Continuous	5,663	49.93	17.29	18.00	94.00
Educational Attainment, Excluded Category = *No High School Degree*					
High School Graduate	5,663	0.30	0.46	0.00	1.00
Some College	5,663	0.20	0.40	0.00	1.00
Associate Degree	5,663	0.09	0.29	0.00	1.00
Bachelors Degree	5,663	0.18	0.39	0.00	1.00
Graduate Degree	5,663	0.13	0.33	0.00	1.00
Race/ Ethnicity, Excluded Category = *White*					
Black	5,662	0.10	0.30	0.00	1.00
Hispanic	5,662	0.10	0.30	0.00	1.00
Other	5,662	0.07	0.25	0.00	1.00
Income, Excluded Category = *Less than $10,000*					
$10,000-$14,999	5,663	0.05	0.22	0.00	1.00
$15,000-$24,999	5,663	0.09	0.28	0.00	1.00
$25,000-$34,999	5,663	0.10	0.30	0.00	1.00
$35,000-$49,000	5,663	0.13	0.33	0.00	1.00
$50,000-$74,999	5,663	0.19	0.39	0.00	1.00
$75,000-$99,999	5,663	0.14	0.34	0.00	1.00
$100,000-$149,000	5,663	0.17	0.37	0.00	1.00

$150,000+	5,663	0.08	0.27	0.00	1.00
% of Respondent's County Which is Foreign Born	5,658	0.12	0.10	0.00	0.51
Urbanization Status of Respondent's County, Excluded Category = *Metro areas of 1 million population or more*					
Metro areas of 250,000 to 1 million population	5,658	0.23	0.42	0.00	1.00
Metro areas of fewer than 250,000 population	5,658	0.10	0.30	0.00	1.00
Nonmetro areas	5,658	0.14	0.35	0.00	1.00
Census Division, Excluded Category = *New England*					
Middle Atlantic	5,658	0.13	0.34	0.00	1.00
Midwest	5,658	0.16	0.37	0.00	1.00
West North Central	5,658	0.08	0.27	0.00	1.00
South Atlantic	5,658	0.20	0.40	0.00	1.00
East South Central	5,658	0.05	0.23	0.00	1.00
West South Central	5,658	0.10	0.30	0.00	1.00
Mountain	5,658	0.07	0.26	0.00	1.00
Pacific	5,658	0.15	0.36	0.00	1.00
Transaction-Level Variables					
Was Transaction Tipped?	1,147	0.91	0.28	0.00	1.00
Tip Rate	924	0.18	0.06	0.01	0.42

Table 10 – Weighted Descriptive Statistics - GfK Sample

Respondent-Level Variables	*N*	Mean	Standard Deviation	Minimum	Maximum
Full Service Restaurant Transactions in Last Day	5,663	0.20	0.45	0.00	4.00
Male	5,663	0.48	0.50	0.00	1.00
Age, Excluded Category = *18-24*					
25-34	5,663	0.19	0.39	0.00	1.00
35-44	5,663	0.17	0.37	0.00	1.00
45-64	5,663	0.36	0.48	0.00	1.00
65+	5,663	0.17	0.38	0.00	1.00
Age, Continuous	5,663	46.87	17.36	18.00	94.00
Educational Attainment, Excluded Category = *No High School Degree*					
High School Graduate	5,663	0.30	0.46	0.00	1.00
Some College	5,663	0.20	0.40	0.00	1.00

Associate Degree	5,663	0.09	0.29	0.00	1.00
Bachelor's Degree	5,663	0.17	0.38	0.00	1.00
Graduate Degree	5,663	0.12	0.32	0.00	1.00
Race/ Ethnicity, Excluded Category = *White*					
Black	5,662	0.11	0.32	0.00	1.00
Hispanic	5,662	0.15	0.36	0.00	1.00
Other	5,662	0.08	0.27	0.00	1.00
Income, Excluded Category = *Less than $10,000*					
$10,000-$14,999	5,663	0.04	0.20	0.00	1.00
$15,000-$24,999	5,663	0.07	0.26	0.00	1.00
$25,000-$34,999	5,663	0.10	0.30	0.00	1.00
$35,000-$49,000	5,663	0.12	0.33	0.00	1.00
$50,000-$74,999	5,663	0.18	0.39	0.00	1.00
$75,000-$99,999	5,663	0.16	0.36	0.00	1.00
$100,000-$149,000	5,663	0.18	0.38	0.00	1.00
$150,000+	5,663	0.08	0.27	0.00	1.00
% of Respondent's County Which is Foreign Born	5,658	0.12	0.10	0.00	0.51
Urbanization Status of Respondent's County, Excluded Category = *Metro areas of 1 million population or more*					
Metro areas of 250,000 to 1 million population	5,658	0.22	0.41	0.00	1.00
Metro areas of fewer than 250,000 population	5,658	0.09	0.28	0.00	1.00
Nonmetro areas	5,658	0.15	0.36	0.00	1.00
Census Division, Excluded Category = *New England*					
Middle Atlantic	5,658	0.14	0.34	0.00	1.00
Midwest	5,658	0.14	0.35	0.00	1.00
West North Central	5,658	0.07	0.26	0.00	1.00
South Atlantic	5,658	0.20	0.40	0.00	1.00
East South Central	5,658	0.06	0.23	0.00	1.00
West South Central	5,658	0.11	0.32	0.00	1.00
Mountain	5,658	0.07	0.26	0.00	1.00
Pacific	5,658	0.16	0.37	0.00	1.00
Transaction-Level Variables					
Was Transaction Tipped?	1,147	0.90	0.30	0.00	1.00
Tip Rate	924	0.18	0.06	0.01	0.42

Table 11 – Unweighted Descriptive Statistics - Ipsos Sample

Respondent-Level Variables	*N*	Mean	Standard Deviation	Minimum	Maximum
Full Service Restaurant Transactions in Last Day	6,920	0.17	0.43	0.00	8.00
Male	6,878	0.46	0.50	0.00	1.00
Age, Excluded Category = *18-24*					
25-34	6,878	0.18	0.39	0.00	1.00
35-44	6,878	0.16	0.36	0.00	1.00
45-64	6,878	0.44	0.50	0.00	1.00
65+	6,878	0.12	0.32	0.00	1.00
Age, Continuous	6,878	46.30	15.78	18.00	105.00
Educational Attainment, Excluded Category = *No High School Degree*					
High School Graduate	6,828	0.21	0.40	0.00	1.00
Some College	6,828	0.26	0.44	0.00	1.00
Associate Degree	6,828	0.12	0.32	0.00	1.00
Bachelor's Degree	6,828	0.25	0.43	0.00	1.00
Graduate Degree	6,828	0.14	0.34	0.00	1.00
Race/ Ethnicity, Excluded Category = *White*					
Black	6,781	0.08	0.26	0.00	1.00
Hispanic	6,781	0.08	0.28	0.00	1.00
Other	6,781	0.08	0.27	0.00	1.00
Income, Excluded Category = *Less than $10,000*					
$10,000-$14,999	6,530	0.06	0.23	0.00	1.00
$15,000-$24,999	6,530	0.12	0.32	0.00	1.00
$25,000-$34,999	6,530	0.11	0.31	0.00	1.00
$35,000-$49,000	6,530	0.14	0.34	0.00	1.00
$50,000-$74,999	6,530	0.19	0.40	0.00	1.00
$75,000-$99,999	6,530	0.12	0.33	0.00	1.00
$100,000-$149,000	6,530	0.12	0.33	0.00	1.00
$150,000+	6,530	0.06	0.24	0.00	1.00
% of Respondent's County Which is Foreign Born	6,914	0.12	0.10	0.00	0.51
Urbanization Status of Respondent's County, Excluded Category = *Metro areas of 1 million population or more*					
Metro areas of 250,000 to 1 million population	6,914	0.22	0.42	0.00	1.00
Metro areas of fewer than 250,000 population	6,914	0.09	0.29	0.00	1.00
Nonmetro areas	6,914	0.13	0.34	0.00	1.00

Census Division, Excluded Category = *New England*					
Middle Atlantic	6,914	0.16	0.36	0.00	1.00
Midwest	6,914	0.18	0.38	0.00	1.00
West North Central	6,914	0.07	0.25	0.00	1.00
South Atlantic	6,914	0.20	0.40	0.00	1.00
East South Central	6,914	0.05	0.22	0.00	1.00
West South Central	6,914	0.08	0.28	0.00	1.00
Mountain	6,914	0.07	0.25	0.00	1.00
Pacific	6,914	0.14	0.35	0.00	1.00
Transaction-Level Variables					
Was Transaction Tipped?	1,144	0.88	0.32	0.00	1.00
Tip Rate	909	0.18	0.06	0.01	0.48

Table 12 – Weighted Descriptive Statistics - Ipsos Sample

Respondent-Level Variables	*N*	Mean	Standard Deviation	Minimum	Maximum
Full Service Restaurant Transactions in Last Day	6,824	0.17	0.44	0.00	8.00
Male	6,824	0.48	0.50	0.00	1.00
Age, Excluded Category = *18-24*					
25-34	6,824	0.18	0.38	0.00	1.00
35-44	6,824	0.15	0.36	0.00	1.00
45-64	6,824	0.44	0.50	0.00	1.00
65+	6,824	0.11	0.31	0.00	1.00
Age, Continuous	6,824	45.74	15.96	18.00	105.00
Educational Attainment, Excluded Category = *No High School Degree*					
High School Graduate	6,824	0.37	0.48	0.00	1.00
Some College	6,824	0.20	0.40	0.00	1.00
Associate Degree	6,824	0.09	0.29	0.00	1.00
Bachelor's Degree	6,824	0.18	0.39	0.00	1.00
Graduate Degree	6,824	0.11	0.31	0.00	1.00
Race/ Ethnicity, Excluded Category = *White*					
Black	6,757	0.11	0.32	0.00	1.00
Hispanic	6,757	0.15	0.35	0.00	1.00
Other	6,757	0.07	0.26	0.00	1.00
Income, Excluded Category = *Less than $10,000*					
$10,000-$14,999	6,530	0.05	0.22	0.00	1.00
$15,000-$24,999	6,530	0.11	0.32	0.00	1.00
$25,000-$34,999	6,530	0.11	0.31	0.00	1.00

$35,000-$49,000	6,530	0.13	0.33	0.00	1.00
$50,000-$74,999	6,530	0.19	0.39	0.00	1.00
$75,000-$99,999	6,530	0.11	0.31	0.00	1.00
$100,000-$149,000	6,530	0.15	0.35	0.00	1.00
$150,000+	6,530	0.07	0.25	0.00	1.00
% of Respondent's County Which is Foreign Born	6,818	0.13	0.11	0.00	0.51
Urbanization Status of Respondent's County, Excluded Category = *Metro areas of 1 million population or more*					
Metro areas of 250,000 to 1 million population	6,818	0.22	0.41	0.00	1.00
Metro areas of fewer than 250,000 population	6,818	0.08	0.28	0.00	1.00
Nonmetro areas	6,818	0.15	0.36	0.00	1.00
Census Division, Excluded Category = *New England*					
Middle Atlantic	6,818	0.14	0.35	0.00	1.00
Midwest	6,818	0.16	0.36	0.00	1.00
West North Central	6,818	0.06	0.23	0.00	1.00
South Atlantic	6,818	0.22	0.41	0.00	1.00
East South Central	6,818	0.06	0.23	0.00	1.00
West South Central	6,818	0.10	0.29	0.00	1.00
Mountain	6,818	0.08	0.27	0.00	1.00
Pacific	6,818	0.16	0.36	0.00	1.00
Transaction-Level Variables					
Was Transaction Tipped?	1,144	0.88	0.32	0.00	1.00
Tip Rate	909	0.18	0.06	0.01	0.48

Analysis

Table 13 – Differences in Samples and Differences in Differences Tests Without Post-Stratification Weights

	Differences in Samples		Differences in Differences			
Variable	Tip Rate	Tip Rate	Difference	Difference	Absolute Difference	Absolute Difference
IPSOS	-0.004 (0.003)	-0.006 (0.003)*	-0.003 (0.003)	-0.005 (0.003)	0.003 (0.002)	0.006 (0.002)*
Male		0.000 (0.003)		0.000 (0.003)		0.004 (0.002)
Age, 25-34		-0.006 (0.008)		-0.007 (0.008)		0.003 (0.005)
Age, 35-44		-0.007 (0.007)		-0.007 (0.008)		-0.003 (0.006)

Age, 45-64	0.004 (0.007)	0.003 (0.007)	-0.006 (0.005)
Age, 65+	0.005 (0.008)	0.004 (0.008)	-0.008 (0.006)
High School Graduate	0.018 (0.010)	0.022 (0.012)	-0.024 (0.008)**
Some College	0.024 (0.009)**	0.026 (0.011)*	-0.028 (0.008)**
Associate Degree	0.025 (0.010)*	0.026 (0.012)*	-0.025 (0.008)**
Bachelor's Degree	0.025 (0.009)**	0.027 (0.011)*	-0.032 (0.008)**
Graduate Degree	0.025 (0.010)**	0.025 (0.011)*	-0.028 (0.008)**
Black	-0.011 (0.007)	-0.011 (0.007)	0.006 (0.004)
Hispanic	-0.017 (0.005)**	-0.016 (0.005)**	0.011 (0.004)**
Other	-0.006 (0.005)	-0.005 (0.005)	0.009 (0.004)*
Income, $10k-$14.9k	-0.000 (0.000)	-0.000 (0.000)	0.000 (0.000)
Income, $15k-$24.9k	-0.000 (0.000)	-0.000 (0.000)	-0.000 (0.000)
Income, $25k-$34.9k	0.000 (0.000)	0.000 (0.000)	-0.000 (0.000)
Income, $35k-$49.9k	0.000 (0.000)	-0.000 (0.000)	-0.000 (0.000)
Income, $50k-$74.9k	0.000 (0.000)	0.000 (0.000)	-0.000 (0.000)
Income, $75k-$99.9k	-0.000 (0.000)	-0.000 (0.000)	-0.000 (0.000)
Income, $100k-$149.9k	0.000 (0.000)	0.000 (0.000)	-0.000 (0.000)
Income, $150k+	0.000 (0.000)	0.000 (0.000)	-0.000 (0.000)*
Foreign Born, % of County Population	0.006 (0.017)	0.056 (0.019)**	-0.037 (0.013)**
Metro Population, 250k - 1 Million	-0.000 (0.004)	0.006 (0.004)	-0.000 (0.003)
Metro Population, <250k	-0.009 (0.005)	0.001 (0.006)	0.004 (0.004)
Non-Metro County	-0.012 (0.005)**	-0.002 (0.006)	0.001 (0.005)
Middle Atlantic	-0.000 (0.006)	-0.008 (0.007)	0.005 (0.004)

Midwest		0.001		-0.010		0.003
		(0.005)		(0.007)		(0.003)
West North Central		-0.015		-0.022		0.012
		(0.007)*		(0.008)**		(0.005)**
South Atlantic		-0.004		-0.025		0.016
		(0.005)		(0.007)**		(0.003)**
East South Central		-0.018		-0.029		0.014
		(0.007)*		(0.012)*		(0.007)
West South Central		-0.012		-0.032		0.025
		(0.006)*		(0.008)**		(0.003)**
Mountain		-0.015		-0.029		0.016
		(0.005)**		(0.008)**		(0.004)**
Pacific		-0.013		-0.007		0.005
		(0.007)		(0.009)		(0.004)
Constant	0.184	0.170	-0.032	-0.044	0.052	0.083
	(0.002)**	(0.018)**	(0.003)**	(0.020)*	(0.002)**	(0.011)**
R^2	.001	.058	.001	.078	.002	.110
N	1,832	1,790	1,723	1,683	1,723	1,683
GfK Predicted Value	0.184	0.185	-0.032	-0.030	0.052	0.051
	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)	(0.001)
Ipsos Predicted Value	0.180	0.179	-0.034	-0.035	0.056	0.057
	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)	(0.002)

Robust standard errors clustered on Commuting Zones in parentheses. Each observation represents a transaction. Column 1 and 2 report results for the unconditional and conditional "Differences in Sample " tests, respectively, where the dependent variable is the transaction. Columns 3 and 4 report results for the unconditional and conditional "Differences in Differences" tests, where the dependent variable is the difference between a transaction's tip rate and the mean tip rate for the respondent's commuting zone derived from the Point of Sale data. Columns 5 and 6 report results for absolute "Differences in Differences" test, where the dependent variable is the absolute difference between a transaction's tip rate and the mean tip rate of the respondent's commuting zone as derived from the Point of Sale Data. The average predicted outcome for the total sample under the counterfactuals that all respondents came from the GfK or Ipsos panels are also presented at the bottom of the table. * p<0.05; ** p<0.01

Table 14 – Differences in Samples and Differences in Differences Tests With Post-stratification Weights

	"Differences in Samples"		"Differences in Differences"			
Variable	Tip Rate	Tip Rate	Difference	Difference	Absolute Difference	Absolute Difference
IPSOS	-0.002	-0.004	-0.001	-0.003	0.003	0.005
	(0.003)	(0.003)	(0.004)	(0.003)	-0.002	(0.002)*
Male		0.000		0.000		0.002
		(0.003)		(0.003)		(0.003)
Age, 25-34		-0.018		-0.019		0.010
		(0.009)*		(0.009)*		(0.005)
Age, 35-44		-0.024		-0.023		0.005
		(0.009)**		(0.009)*		(0.006)

Age, 45-64	-0.011 (0.008)	-0.013 (0.008)	0.000 (0.005)
Age, 65+	-0.008 (0.009)	-0.009 (0.010)	-0.002 (0.006)
High School Graduate	0.023 (0.010)*	0.026 (0.013)*	-0.030 (0.009)**
Some College	0.027 (0.010)**	0.029 (0.012)*	-0.030 (0.009)**
Associate Degree	0.030 (0.011)**	0.032 (0.013)*	-0.030 (0.009)**
Bachelor's Degree	0.024 (0.010)*	0.026 (0.012)*	-0.035 (0.009)**
Graduate Degree	0.028 (0.011)**	0.027 (0.013)*	-0.035 (0.009)**
Black	-0.007 (0.007)	-0.007 (0.007)	0.004 (0.004)
Hispanic	-0.014 (0.005)**	-0.013 (0.006)*	0.009 (0.004)*
Other	-0.012 (0.005)*	-0.011 (0.005)*	0.007 (0.004)
Income, $10k-$14.9k	-0.000 (0.000)	-0.000 (0.000)	-0.000 (0.000)
Income, $15k-$24.9k	-0.000 (0.000)	-0.000 (0.000)	-0.000 (0.000)
Income, $25k-$34.9k	-0.000 (0.000)	-0.000 (0.000)	-0.000 (0.000)
Income, $35k-$49.9k	-0.000 (0.000)	-0.000 (0.000)	-0.000 (0.000)
Income, $50k-$74.9k	0.000 (0.000)	-0.000 (0.000)	-0.000 (0.000)
Income, $75k-$99.9k	-0.000 (0.000)	0.000 (0.000)	-0.000 (0.000)
Income, $100k-$149.9k	0.000 (0.000)	0.000 (0.000)	-0.000 (0.000)*
Income, $150k+	0.000 (0.000)	0.000 (0.000)	-0.000 (0.000)*
Foreign Born, % of County Population	0.020 (0.019)	0.072 (0.022)**	-0.044 (0.015)**
Metro Population, 250k - 1 Million	0.006 (0.005)	0.013 (0.005)*	-0.002 (0.004)
Metro Population, <250k	-0.006 (0.005)	0.003 (0.006)	0.004 (0.004)
Non-Metro County	-0.007 (0.005)	0.002 (0.007)	0.000 (0.006)
Middle Atlantic	0.003 (0.007)	-0.004 (0.008)	0.007 (0.004)

Midwest		0.005 (0.006)		-0.006 (0.007)		0.003 (0.003)
West North Central		-0.017 (0.008)*		-0.025 (0.009)**		0.015 (0.005)**
South Atlantic		-0.002 (0.006)		-0.024 (0.007)**		0.016 (0.004)**
East South Central		-0.016 (0.008)		-0.027 (0.012)*		0.015 (0.009)
West South Central		-0.006 (0.006)		-0.026 (0.008)**		0.024 (0.004)**
Mountain		-0.013 (0.006)*		-0.029 (0.009)**		0.017 (0.005)**
Pacific		-0.006 (0.007)		0.000 (0.009)		0.005 (0.004)
Constant	0.180 (0.002)**	0.176 (0.027)**	-0.035 (0.003)**	-0.039 (0.029)	0.055 (0.002)**	0.097 (0.014)**
R^2	.000	.067	.000	.099	.001	.122
N	1,832	1,790	1,723	1,683	1,723	1,683
GfK Predicted Value	0.180 (0.002)	0.181 (0.002)	-0.035 (0.003)	-0.033 (0.002)	0.055 (0.002)	0.054 (0.002)
Ipsos Predicted Value	0.179 (0.003)	0.177 (0.003)	-0.036 (0.004)	-0.037 (0.003)	0.058 (0.002)	0.059 (0.002)

Robust standard errors clustered on Commuting Zones in parentheses. Each observation represents a transaction. Column 1 and 2 report results for the unconditional and conditional "Differences in Sample Tests", respectively, where the dependent variable is the transaction. Columns 3 and 4 report results for the unconditional and conditional "Differences in Differences" tests, where the dependent variable is the difference between a transaction's tip rate and the mean tip rate for the respondent's commuting zone derived from the Point of Sale data. Columns 5 and 6 report results for absolute "Differences in Differences" test, where the dependent variable is the absolute difference between a transaction's tip rate and the mean tip rate of the respondent's commuting zone as derived from the Point of Sale Data. Observations are weighted using normalized post -stratification weights provided by Ipsos and GfK. These weights were normalized to 1 for each sample and then divided by 2 so that the combined sample weights sum to 1. The average predicted outcome for the total sample under the counterfactuals that all respondents came from the GfK or Ipsos panels are also presented at the bottom of the table. * $p<0.05$; ** $p<0.01$

Table 15 – Determinants of Membership in the Ipsos Sample

	Linear Regression		Logit Regression	
Variable	Unweighted	Weighted	Unweighted	Weighted
Male	-.017 (.010)	.004 (.011)	-0.075 (0.044)	0.017 (0.047)
Age, 25-34	-.036 (.020)	-.034 (.022)	-0.166 (0.086)	-0.137 (0.092)
Age, 35-44	-.042 (.019)*	-.029 (.021)	-0.190 (0.083)*	-0.120 (0.091)
Age, 45-64	-.030 (.016)	.029 (.018)	-0.137 (0.070)	0.125 (0.078)
Age, 65+	-.217 (.018)**	-.137 (.021)**	-0.957 (0.079)**	-0.582 (0.089)**

High School Graduate	.212	.281	1.017	1.259
	(.017)**	(.019)**	(0.093)**	(0.100)**
Some College	.376	.252	1.719	1.137
	(.019)**	(.023)**	(0.106)**	(0.115)**
Associate Degree	.384	.256	1.758	1.157
	(.021)**	(.024)**	(0.108)**	(0.119)**
Bachelor's Degree	.432	.299	1.973	1.337
	(.019)**	(.022)**	(0.104)**	(0.110)**
Graduate Degree	.421	.289	1.926	1.296
	(.021)**	(.025)**	(0.112)**	(0.121)**
Black	-.119	-.051	-0.528	-0.217
	(.016)**	(.018)**	(0.072)**	(0.079)**
Hispanic	-.065	-.010	-0.289	-0.043
	(.016)**	(.017)	(0.071)**	(0.074)
Other	-.018	-.029	-0.083	-0.125
	(.023)	(.034)	(0.100)	(0.142)
Income, $10k-$14.9k	-.001	-.001	-0.003	-0.003
	(.000)**	(.000)*	(0.001)**	(0.001)*
Income, $15k-$24.9k	.000	.000	-0.001	0.001
	(.000)	(.000)	(0.001)	(0.001)
Income, $25k-$34.9k	-.001	-.001	-0.004	-0.003
	(.000)**	(.000)**	(0.001)**	(0.001)**
Income, $35k-$49.9k	-.001	-.001	-0.004	-0.004
	(.000)**	(.000)**	(0.001)**	(0.001)**
Income, $50k-$74.9k	-.001	-.001	-0.006	-0.005
	(.000)**	(.000)**	(0.001)**	(0.001)**
Income, $75k-$99.9k	-.002	-.002	-0.009	-0.010
	(.000)**	(.000)**	(0.001)**	(0.001)**
Income, $100k-$149.9k	-.003	-.002	-0.012	-0.008
	(.000)**	(.000)**	(0.001)**	(0.001)**
Income, $150k+	-.003	-.002	-0.013	-0.008
	(.000)**	(.000)**	(0.001)**	(0.001)**
Foreign Born, % of County Population	.160	.129	0.704	0.541
	(.056)**	(.063)*	(0.252)**	(0.263)*
Metro Population, 250k - 1 Million	-.019	-.018	-0.084	-0.076
	(.012)	(.014)	(0.054)	(0.060)
Metro Population, <250k	-.016	-.022	-0.070	-0.094
	(.017)	(.020)	(0.075)	(0.083)
Non-Metro County	-.024	-.022	-0.106	-0.093
	(.016)	(.018)	(0.070)	(0.075)
Middle Atlantic	.039	.021	0.172	0.091
	(.023)	(.027)	(0.102)	(0.114)
Midwest	.028	.047	0.121	0.198
	(.024)	(.029)	(0.106)	(0.123)
West North Central	-.028	-.017	-0.127	-0.079
	(.029)	(.035)	(0.127)	(0.149)
South Atlantic	.022	.048	0.099	0.202
	(.022)	(.027)	(0.097)	(0.113)
East South Central	.009	.022	0.044	0.092
	(.026)	(.030)	(0.112)	(0.125)
West South	-.018	-.003	-0.079	-0.017

Central	(.023)	(.030)	(0.101)	(0.124)
Mountain	-.011	.038	-0.050	0.158
	(.023)	(.029)	(0.102)	(0.123)
Pacific	-.040	.005	-0.176	0.023
	(.025)	(.031)	(0.109)	(0.130)
Constant	.407	.341	-0.468	-0.734
	(.034)**	(.039)**	(0.161)**	(0.173)**
R^2	.092	.054	0.070	0.040
N	12,137	12,137	12,137	12,137

Robust standard errors clustered on Commuting Zones in parentheses. Each observation represents a respondent. The dependent variable in all cases is a dichotomous variable that takes a value of 1 if the respondent is a member of the Ipsos sample and 0 if the respondent is a member of the GfK knowledge panel. Column 1 and 2 report unweighted and weighted results for a linear probability model, respectively. Columns 3 and 4 reports mean marginal effects for each variable derived from a logit models of sample membership. Post -stratification weights were normalized to 1 for each sample and then divided by 2 so that the combined sample weights sum to 1. * $p<0.05$; ** $p<0.01$

Table 16 – Unconditional Tests Excluding Observations With Missing Data on Control Variables

Unweighted			
	Tip Rate	Difference	Absolute Difference
IPSOS	-0.004	-0.003	0.004
	(0.003)	(0.003)	(0.002)
Constant	0.184	-0.032	0.052
	(0.002)**	(0.003)**	(0.002)**
R^2	0.001	0.001	0.002
N	1,790	1,693	1,693
Weighted			
	Tip Rate	Difference	Absolute Difference
IPSOS	-0.002	-0.001	0.003
	(0.003)	(0.004)	(0.002)
Constant	0.180	-0.035	0.055
	(0.002)**	(0.003)**	(0.002)**
R^2	0.000	0.000	0.001
N	1,790	1,693	1,693

FORS|MARSH
GROUP

Interim Report on the Survey of Consumer Tipping Behavior

Prepared for the Internal Revenue Service

Prepared by Fors Marsh Group, LLC

October 2017

Version 2 revised January 2018

TABLE OF CONTENTS

Introduction

There is limited nationally representative information concerning the amount of tipped income relative to total national income. The private nature of most transactions that involve tipping makes it extremely difficult to collect reliable data that can be used to estimate total tip income. This difficulty is further compounded by the motivation of some individuals not to report their tips as taxable income. For these reasons, surveying consumers about their tipping experiences may be the most reliable way to collect quantitative data on tip income. However, the last large-scale survey of consumer tipping behavior was undertaken by the Internal Revenue Service (IRS) in 1982. To provide updated estimates of consumer tipping behavior, the IRS has begun conducting a yearlong survey of consumer tipping behavior—from here on referred to as the "Full Fielding"—over the course of calendar year 2017. As of the publication of this report, data have been collected for the first six months of the year (i.e., January through June of 2017). This report presents preliminary estimates of tipping prevalence and tipping rates by both industry and geography for those first six months.

Survey Administration

The target population for the IRS tipping study includes all U.S. residents who use services that are commonly tipped. The number of individuals in this population is unknown, but likely includes a majority of the U.S. adult population. Example settings where tipping is typical include full-service restaurants, taxis, barbershops, beauty salons, hotels, and casinos.

Prior IRS research on consumer tipping behavior found tipping rates varied considerably by industry and by region. The previously mentioned 1982 study was conducted by the University of Illinois for the IRS[1] and found that tipping rates were roughly 14% of the total bill for restaurants, 12% for barber and beauty shops, 19% for bars, and 20% for taxis. On a regional basis, mean restaurant tipping rates ranged from a low of 12.5% in the West North Central region to a high of 15% in the Northeast.

The observed variation in tipping rates implies that larger sample sizes are required to produce accurate estimates of tipping rates. All else being equal, a larger sample size means greater cost. This constraint may be met in two ways: (1) limiting the scope of the study to focus on fewer industries or regions, or (2) finding a more cost-effective mode of data collection. Because of the previous study's finding on the variance of tipping rates by industry and region, the IRS believes it would be inappropriate to limit the scope of this study, and therefore decided to pursue a lower cost mode of data collection.

With respect to lowering the cost of data collection, an increasingly common alternative is the use of nonprobability internet samples. The costs of sampling from an opt-in internet panel may be substantially lower than the costs associated with sampling from a telephone- or mail-based frame, or a panel recruited from such frames (e.g., probability-based web panel). In addition, there might be additional costs or nonresponse associated with pushing individuals sampled from the telephone or mail frame to the internet survey instrument. The chief drawback of using a nonprobability sample from an opt-in internet panel is that such panels could produce a realized sample that is less representative of the target population than the phone or mail frames. However, given the high rates

[1] Pearl, R. B., & Sudman, S. (1983, June). *A survey approach to estimating the tipping practices of consumers* (Final Report to the Internal Revenue Service under Contract TIR 81-52); Pearl, R. B. (1985, July). *Tipping practices of American households: 1984* (Final Report to the Internal Revenue Service under Contract 82-21).

of nonresponse associated with sampling from phone or mail frames, it is not clear to what degree respondents from probability samples are more representative with respect to tipping behavior than respondents contacted through an opt-in internet panel, particularly after poststratifying on observed demographic characteristics. Although nonresponse can be mitigated through follow-up contacts, doing so can exacerbate the differences between the probability and nonprobability sampling strategies with respect to the cost of obtaining a sample of a given size. Consequently, given a fixed budget, it is unclear whether the reductions in bias in the estimates of mean tipping and stiffing rates that result from using a probability sample is worth the increase in the variability of these estimates that results from a smaller sample size, especially for relatively infrequent tipping transactions. A pilot study undertaken by the IRS comparing estimates derived from both a probability and nonprobability sample generally failed to find statistically significant differences between the two samples.

The nonprobability sample was chosen for this current study and is collected from Ipsos' blended panel. Ipsos' blended sample approach combines the use of its Ampario online sampling method in addition to its i-Say online panel—an online panel of 800,000 members and their households. Ampario is a nonprobability sampling procedure developed by Ipsos that invites respondents by invitations, banner ads, and other means on 100 to 400 websites that have partnered with Ipsos. These two methods are combined into a single sample using Ipsos' proprietary Cortex routing system, which allocates and reallocates a sample based on respondent eligibility. Simply put, when respondents are not eligible for one survey, they are immediately redirected to other surveys in progress. In traditional one-off, opt-in surveys, noneligible respondents are lost, representing a considerable cost. Finally, Bayesian methodology, which requires previous information regarding the overall sample of interest in order to mix with current information for the final distribution of results, is used to form the final distribution. As is the case with a traditional online sample, Ipsos' blended sampling could work with several different data collection modes, but it is best served with an online-based questionnaire, which could include a cross-sectional administration or a longitudinal diary approach. However, because of the opt-in nature of the blended sample, it is not possible to model the probability of responding to a survey, thus there exists that source of potential bias in survey estimates.

Recruitment Sources Used in the Project

Ipsos i-Say	Ipsos panels are not just lists or databases of individuals, but actively managed research Access Panels: • Individuals who have volunteered to take part in market research surveys • Created and managed for long-term use and access • Extensively profiled to efficiently target respondents The vast majority of panelists are referred to Ipsos through various online suppliers. Ipsos only uses high-quality recruitment sources to entice people who are eager to take surveys. The organization strategically focuses on developing processes that reflect the newest internet practices as may currently be found through social networks. Email lists, banners, website and text ads, co-registration, and search engine marketing are also used.
Lightspeed GMI	This is an actively managed panel composed of people who made a conscious decision to participate in online surveys through a double opt-in registration

	process. Several methodologies are used to recruit panelists, including opt -in email, co-registration, e-newsletter campaigns, and traditional banner placements, as well as both internal and external affiliate networks. Social media is included through Lightspeed's recruiting partners.
Market Cube	Market Cube owns and operates the Univox Community , an actively managed panel with an individual -level compensation model. Market Cube also has access to a vast network of social media and publisher respondents that can be used to supplement internal assets. Additionally, Market Cube has developed close relationships with a variety of panel companies with which they can partner on difficult -to-reach subpopulations. These strategic partnerships allow Market Cube to leverage relevant lists, databases, and networks to fulfill specific client requirements.
ROIRocket	This large ad network has provided more than 30 million panelists to date and offers access to more than 5 million active respondents at any given time. The company has experience in using its sample for online communities, custom panels, in-depth interviews, longitudinal research studies, etc.
SSI	This is an actively managed panel incorporating participants from partnership sources managed by SSI, recruited via banners, invitations , and messaging. Prospects go through rigorous quality controls before being included in SSI panels.

Quota Sampling Methods and Variables

Sample Balancing

Ipsos and each of its partners select ed what is known as a "balanced return" sample, wherein the demographic distribution of "clicks" (meaning respondents who respond to a survey invitation by clicking the hyperlink and entering the survey) match es the demographic distribution of the overall U.S. population, as indicated in most recent results of the Census Bureau's Current Population Survey (CPS).[2] Because different individuals and demographic groups respond at different rates, the different samp ling rates are applied for these different groups. The demographic distribution of the contacted sample , thus, does not match the demographic distribution of the U.S. population.

Sample balancing (i.e. , determining the proportion of the sample to allocate to different demographic groups) was done using four demographic variables: gender, age, region, and income. The links between each of these characteristics and tip rates have been the subject of past academic studies on tipping behavior . These variables were fully crossed, creating 96 sampling cells (see Table A1). The levels (sample groups) within each of the variables are indicated in Table 1 .

[2] To ensure sufficient sample records to complete the necessary number of interviews each month, multiple sample sources are needed. The sample for the IRS Consumer Tipping Study is provided by Ipsos' opt -in i-Say panel and four other opt-in panels, with the anticipated proportion of completed interviews provided by each source remaining constant each month (and following the proportions used in the pilot test). Each panel provider ha s prepared responses to ESOMAR's 28 questions for online samples and has been vetted by Ipsos' online research department. These panel providers will email invitations to their panelists with a link that directs them to the Ipsos survey site after passing them through an intermediary site used by the panel provider to monitor whether they (A) respond and (B) complete the survey, so that their traditional panel incentive is paid. Panel partners will provide information on how many invitations are sent and wi ll balance their samples using targets provided by Ipsos.

Table 1: Stratification Variables

Gender	Age	Region	Income
(1) Male	Age 18–34	(1) Northeast	Under $20K
(2) Female	Age 35–54	(2) Midwest	$20K–$49,999
	Age 55+	(3) South	$50K–$99,999
		(4) West	$100K+

Ipsos selected samples two times a week (Monday and Friday). On Monday, the sample was designed to produce a demographically balanced return sample equal to four days' total of completed interviews. On Friday, the sample was designed to produce the balanced return sample equal to three days' total of completed interviews. The samples were divided into replicates or subsamples that equally represent the larger sample (four replicates for the Monday samples ; three replicates for the Friday samples), so that one replicate could be "released" (meaning survey invitations were sent to those sampled individuals) each day. These invitations, which included invitation text, a link to the survey program, and a link to the panel provider's member policies (including confidentiality), followed the standard email invitation formats used by Ipsos and each of its partners, so that sampled individuals were familiar with how to use them to access the survey. This approach yielded approximately the targeted 144 daily completed interviews.

This approach of using sample replicates is employed to achieve greater efficiency when many sample balancing cells are employed by ensuring higher response rates in relatively sparse sampling cells.

The sample design assumed a one-month reuse of sample (i.e. , individuals who were sampled for the study in one month were ineligible for another contact until the next month).

Quality Assurance Processes

Data Collection and Sample Quality and Security Procedures

Ipsos employed a number of quality checks during the data collection process.

- Survey level:
  - Filtering of respondents based on participation history
  - Respondent screening based on demographic variables being captured for the survey (age, gender, ZIP code, etc.)

- Engine level:
  - GeoIP verification : validates survey country versus respondent country determined based on IP
  - Language verification : validates survey language versus respondent language
  - Device check : match between device used by respondent and the device setting of the survey
  - Algorithm to identify possibly unengaged respondents (straight -lining, speeding, providing invalid verbatim in open -ended questions)
  - Concurrent session sniffout : filter respondents with more than one opened session, in the same browser, on the same survey

- Fraud Profile Flag 4 (FPF4) : machine time versus time based on geolocation mismatch
- Open and anonymous proxy checks
- VOID: analysis of web cookies, PanelistID/SupplierID (identifiers provided by sample sources), RelevantID (third-party security service), SHA-1 hash function

Data Analysis Quality Assurance Procedures

Web Survey Quality Control. The FMG Team performed full testing of the programmed instrument to ensure that skip logic, randomization, conditional data piping, question wording, and all other specifications for the survey instrument were met. FMG's online survey quality control process was thorough and included checks to ensure that there were no grammatical or formatting errors, that the question type was accurate (single punch vs. multi punch, etc.), that skip patterns functioned appropriately, and that data restrictions for open-ended questions matched requirements. The FMG Team also had data capture checks in place to examine the functionality of the programmed survey. As a standard quality control check, multiple FMG researchers responded to the online survey and simultaneously recorded the answers on a paper copy of the survey; during these checks, researchers tested all branching/paths of skip patterns in the questionnaire.

Survey Tracking. We established and maintained a secure survey control system that documented the correspondence and tracked the status of all sample members. The heart of this system is a unique sample ID that was given to each sample member and used in place of name, address, or other personally identifiable information. All correspondence—including any emails, phone calls, or other correspondence with a respondent—was logged and coded with a disposition based on the reason for the contact. This process ensured that all sample members were accounted for and given the proper disposition code in line with American Association for Public Opinion Research (AAPOR) and Council of American Survey Research Organizations (CASRO) guidelines. This ultimately allowed the FMG Team to appropriately calculate cooperation and response rates and track issues and problems with the survey effort.

Data Verification and Cleaning. Once data collection was completed and all survey data entered, the data sets were reviewed and thoroughly checked before any analyses were conducted. Records were inspected to determine whether any completed cases should have been discarded. These data quality control checks were made to ensure that the analysis file was clean. Table 2 below details the minimum steps taken.

Table 2: Data Cleaning Steps

Data Cleaning Steps Taken Prior to Analysis	
1) Receive data sets	9) Check skip patterns
2) Print format library (file information)	10) Check recodes
3) Run frequencies (weighted and unweighted)	11) Check calculated variables
4) Check variable names	12) Check coding of 'other, specify'
5) Check variable labels	13) Address problems
6) Check value labels	14) Make changes to formats
7) Check weights (against known pop . totals)	15) Secondary review of final data set
8) Check unweighted sampling	16) Recheck all resultant values

Data Cleaning

Mismeasurement in survey responses can bias estimated stiffing and tipping rates. To mitigate bias, several data cleaning procedures were applied to the survey data.

Repeat Respondents

Although individuals were prevented from responding to the survey multiple times in a given month, there was no procedure in place to prevent individuals from responding and repeating the survey in different months. Because a given individual's tipping behavior ove r time may be more similar than the tipping behavior of two different individuals over time, the responses of repeat respondents across survey completes may not be independent, which can complicate statistical inference.

In addition, there is some eviden ce from prior research on consumer panel surveys that exposure to the survey instrument or the completion of a survey may influence respondent spending and saving behavior.[3] Consequently, individuals who have already responded to a survey may no longer be representative of the wider population of interest with respect to tipping behavior.

Table 3: Number of Respondents by Number of Completed Surveys

Number of Completed Surveys	Number of Respondents
1	20,266
2	908
3	174
4	51
5	1
Total	21,400

To mitigate these issues, for individuals who completed the survey multiple times, only data from the first completed survey were retained for analysis. Of the 21,400 respondents, 1,134 (5.3%) had more than one completed survey (Table 3). A total of 1,413 out of 22,813[4] total completes (6.2% of the total) were dropped as a result of this procedure.

[3] Crossley, T. F., Bresser, J., Delaney, L., & Winter, J. (2017). Can survey participation alter household saving behaviour?. *The Economic Journal*.
[4] This number excludes 156 completes that were classified as Abandoned, Error, or Quota Full Client.

Repeated Transactions

There was also evidence that within a given completed survey, respondents were reporting the same transaction multiple times. Specifically, duplicate transactions for a given respondent-day were identified based on tipping industry, bill size, and tip amount. These duplicate transactions may reflect confusion on the part of the respondent with respect to survey instructions, whereby respondents may be unsure about the requested recall period, and thus may be recording transactions that took place over multiple days. Alternatively, respondents may have been confused as to whether the information about a given transaction was actually recorded, and thus decided to enter it again.

To mitigate potential bias that results from these duplicate entries, for a set of transactions reported by a given respondent for a given day with the same subindustry, bill size, and tip amount, only one transaction was retained. A total of 1,095 out of 32,173 total transactions (3.4% of total) were dropped as a result of this deduplication.

Detection of Extreme Values

We observed several instances of extremely high bill amounts, tip amounts, and tip rates in the survey data. Assuming some of the unusual and unexpected data points represent measurement error or invalid transactions, an outlier identification strategy similar to that employed by FMG in the IRS Tipping Task Order 3 report, *Comparison of Estimates of Tipping Behavior Produced Using Probability and Non-Probability Samples: Methodology and Results*, were implemented for the current study.

Specifically, we assumed that total daily expenditure on bills and tip amounts are log normally distributed and tip rate is normally distributed for each transaction type (e.g., full-service restaurants, hairdressers).[5][6] Total bill and tip expenditure were used to identify outliers rather than characteristics of individual transactions because the expenditures combine information on transaction frequency and transaction characteristics, both of which are necessary for calculating total tipped expenditure or transaction weighted stiffing and tipping weights. We then calculated the following ratio for each outcome by transaction type as follows:

$$\frac{|x - Q_{75}|}{Q_{75} - Q_{25}} \text{ for } x > Q_{75}$$

$$\frac{|x - Q_{25}|}{Q_{75} - Q_{25}} \text{ for } x < Q_{25}$$

---

[5] We recognize the normality assumption applied may not hold due to non-independence of transactions within commuting zones as well as individual respondents. However, the small number of transactions per commuting zone and individual makes identifying outliers by commuting zone and individual unfeasible. It should also be noted that the standard errors do not account for the identification of outliers. Under a different sample, the threshold for identifying outliers would be different, resulting in potentially significantly different estimates. This, along with uncertainty surrounding missing data for certain transaction characteristics, means that the resulting point and standard error estimates could be sensitive to minor changes in methodology, particularly for industries with smaller numbers of transactions.

[6] Results dropping only observations that were identified as outliers with respect to activity at full-service restaurants are presented in Appendix C. The estimates for full-service restaurants are very close to the baseline estimates, but the estimates for all other transactions are severely affected by the inclusion of outliers.

In this case, *y* is logged total daily bill on a given service, logged total daily tips given to a type of service, or the ratio of total daily tips over total daily bills. Respondents are identified as outliers if, for a given transaction type, either of the above ratios exceeds 2.5 for bill amount, tip amount, or tip-to-bill ratio.[7] Respondents with at least one outlier transaction type are excluded from the analysis. A total of 3,241 out of 21,400 remaining respondents (15.1%) were identified as outliers using this procedure. The sample remaining after these exclusions consists of 18,159 respondents.

Estimation Procedures

Given survey nonresponse as well as systematic differences between those respondents dropped from the survey due to a suspected high degree of measurement error in their responses, the final set of respondents may not be representative of the population with respect to characteristics relevant to tipping behavior. This lack of representation could in turn result in biased estimates of average tip rates and stiffing rates. To mitigate such bias, two forms of poststratification are employed to make the estimates reflect the tipping behavior of the general adult population rather than simply the estimation sample: poststratification weights and Multilevel Regression and Poststratification (MRP).

Poststratification Weights

When calculating estimates of transaction tipping or stiffing rates based on a sample, simple transaction averages may be biased. This potential for bias is because the sample is not representative of the population with respect to characteristics relevant to tipping behavior. Poststratification weights are used in such circumstances to calculate weighted averages, in which greater weight is given to respondents whose characteristics are underrepresented in the sample relative to the population of interest, and which in turn reduces estimate bias.

To calculate poststratification weights, a simple raking algorithm was used. Initially, each respondent is given equal weights (i.e., values of 1). The algorithm starts by comparing the distribution of respondents across categories of one characteristic, such as age, to the distribution of the target population. Respondents' initial weights are adjusted by multiplying it by the ratio of the fraction of the population in the respondent's category to the fraction of the sample in the respondent's category. The process is then repeated for another variable, but using the adjusted weights from the previous weights to calculate the fraction of the sample in a given category and, thus, the next adjustment. The process is replicated for all relevant variables, and then another cycle through each variable is initiated using the adjusted weights from the previous cycle. The raking algorithm ensures that the final weighted distribution of the variables used to rake in the sample is very close to those in the population. Finally, the weights are scaled such that they sum to the product of the population of individuals 18+ in 2015[8] by 365 days to facilitate the calculation of estimates of annual totals of tipped expenditure.

[7] For a given variable, a 2.5 interquartile range (IQR) threshold would only identify approximately 0.005% of respondents as outliers under a normal distribution, and is thus a relatively conservative threshold.

[8] The year 2015 is used because it is the last year for which the 5-year American Community Survey (ACS) estimates of the population were available.

The set of variables used for raking, along with weighted and unweighted sample proportions, are presented in Table A2. Note that 127 respondents were dropped due to lack of data for at least one of these post-stratification variables, leading to a final estimation sample of 18,032 respondents.

Multilevel Regression and Poststratification (MRP)

The IRS intends to use the consumer tipping data from this survey in a number of ways. One of those ways will be to develop subnational, industry-specific tipping rates. This section provides a discussion of how FMG developed those rates from the survey data.

One means of obtaining both nationally and subnationally representative estimates of tipping and stiffing rates is MRP (Gelman & Little, 1997;[9] see Buttice and Highton, 2013,[10] and Toshkov, 2015,[11] for recent reviews and critiques). Model-based poststratification strategies have been employed to generate estimates that conform to administrative data using non-representative samples.[12] MRP has attained popularity among social scientists who wish to obtain geographically disaggregated estimates of a quantity of interest. Awareness of variation in tipping rates faced by establishments in different parts of the country will be of potential use to the IRS in so far as it provides a general understanding of patterns of tipping behavior and it might help detect differences in compliance.

Analyzing consumer tipping data for a particular industry using MRP involves estimating models of the number of transactions undertaken by consumers as well as their tipping behavior that take the form:

$$\hat{\square}_{ik}\square = \square^{\beta\square X_{ik} + \alpha\square G_k + C_{\square\square}}$$

$$\hat{\square}_{tik}\square = \frac{\square^{\beta\square X_{ik} + \alpha\square G_k + C_S}}{1 + \square^{\beta\square X_{ik} + \alpha\square G_k + C_S}}$$

$$\hat{\square}_{tik}|(\square_{tik}\square = 0) = \square^{\beta\square X_{ik} + \alpha\square G_k + C_{\square\square}}$$

$$\hat{\square}_{tik}|(\square_{tik}\square = 1) = \square^{\beta\square X_{ik} + \alpha\square G_k + C_{\square\square}}$$

$$\hat{T}_{tik} = \square^{\beta\square X_{ik} + \alpha\square G_k + C_T}$$

...in which $\hat{\square}_{ijk}\square$ is the expected total number of transactions engaged in by respondent *i* in location *k*; $\hat{\square}_{tik}\square$ is the expected probability that respondent's transaction *t* was tipped; $\hat{\square}_{tik}\square$ is the expected bill size for respondent's transaction *t*, which is allowed to vary based on whether or not it was tipped;[13]

---

[9] Gelman, A., & Little, T. C. (1997). Post-stratification into many categories using hierarchical logistic regression. *Survey Methodology, 23* (2): 127–135.

[10] Buttice, M. K., & Highton, B. (2013). How does multilevel regression and post-stratification perform with conventional national surveys?. *Political Analysis*, 21(4), 449–467.

[11] Toshkov, D. (2015). Exploring the performance of multilevel modeling and post-stratification with eurobarometer data. *Political Analysis*, mpv009.

[12] Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting*, *31*(3), 980–991.
Goel, S., Obeng, A., & Rothschild, D. Non-representative surveys: Fast, cheap, and mostly accurate. Working Paper.

[13] The exception for this is full-service restaurants, for which only one average bill size is calculated, due to the small fraction of transactions that were not tipped.

and $\hat{T}_{tik}$ is an expected tip rate for transaction *t* calculated by dividing a reported dollar amount in tips by transaction bill size; *X* is a set of observable respondent-level demographic variables that includes age, gender, and educational attainment, and that are likely to be correlated with both tipping behavior and the number of transactions; and *G* is a set of location-specific factors that include: the racial composition of the respondent's county (i.e., percentage Black, Hispanic, and Other); the percentage of the adult population that is foreign born; the fraction of households in the respondent's county in a given income bracket/ median household income of the county; size of respondent's metropolitan area/ whether the respondent is residing within a metropolitan area; and census region. These variables are intended to capture variability in the number of transactions and tipping behavior by sector that is not explained by differences in *X* between locations. See Table A3 for sample and population proportions for all predictors. Note that while the location *k* is the most narrowly defined geographic area for which data is available, predictions can be generated for aggregated levels of geography *g*. Finally, *C* is a constant.

Parameters β, α, and *C*, and predictions of $\hat{\square}$, $\hat{\square}$, and $\hat{\square}$ are estimated via Poisson regression, whereas parameters for $\square$ are estimated using a logistic regression. The resulting models are used to generate predictions for each outcome for each strata defined by all *N* combinations of values of *X* and *G* covariates. Poststratification is then used to generate the predicted annual number of tipped transactions, transaction average stiffing rates, tipping rates, and ratios of total tipped expenditure to total bill size for a given location:

$$\widehat{\# \square} = \sum_{s}^{N} \hat{E}_s \hat{\square}_s P_s \, 365$$

$$\overline{(1 - \hat{\square})}_{\square} = \sum_{s}^{N} \frac{\hat{E}_s P_s}{\sum_s^N \hat{E}_s P_s} (1 - \hat{\square}_{tik})$$

$$\bar{\square}_{\square} = \sum_{s}^{N} \frac{\hat{E}_s \hat{\square}_s P_s}{\sum_s^N \hat{E}_s \hat{\square}_s P_s} \hat{T}_s$$

$$\overline{\widehat{\square}}_{\square} = \frac{\overline{\widehat{\square}}_{\square}}{\overline{\widehat{\square}}_{\square}}, \text{ where:}$$

$$\overline{\widehat{\square}}_{\square} = \sum_{s}^{N} \hat{E}_s \hat{\square}_s P_s \hat{\square}_{tik} | (\square_{tik} = 1) \, \hat{T}_s$$

$$\overline{\widehat{\square}}_{\square} = \sum_{s}^{N} (\hat{E}_s \hat{\square}_s P_s \hat{\square}_{tik} | (\square_{tik} = 1)) + (\hat{E}_s (\hat{1} - \square_s) P_s \hat{\square}_{tik} | (\square_{tik} = 0))$$

P is the population of a given demographic/geographic stratum *s* in a given location *g*, taken from the 2015 five-year American Community Survey (ACS). Commuting zone-level geographic factors are used to model individuals' number of transactions and tipping behavior. Predictions are generated for the United States as a whole as well as for commuting zones. The preferred subnational geographic unit is the commuting zone. Commuting zones are more likely to encompass the customer base of a given establishment. Commuting zones have been used in recent, prominent studies to define the geographic extent of environmental determinants of social outcomes.[14] Commuting zones may act as a proxy for the typical geographic extent of respondents' daily travels, and thus the establishments they are likely to visit.

This MRP procedure was undertaken separately for each industry, excluding home maintenance, hotels, and casinos, where either there is a significant likelihood that the transaction took place outside the respondent's commuting zone of residence or the number of transactions is extremely low. To quantify the uncertainty in the estimates that results from sampling variability, a cluster bootstrap procedure was used. Specifically, 1,000 samples of commuting zones were drawn with replacement (i.e. the sample commuting zone can enter multiple samples), and data from respondents from a given replicate sample of commuting zones were used to generate the MRP estimates. This resulted in 1,000 replicate estimates of the transaction average stiffing rate, tipping rate, and tipping percentage for each county or commuting zones. The standard deviation of the replications is the standard error of the estimate. A separate table includes estimates and standard errors for each commuting zone.

Results

This section describes both national and commuting zone estimates of tipping and stiffing rates. Definitions of terms used in the section are presented in Table 4.

Table 4: Glossary of Terms

Outcome	Definition
Number of Daily Transactions	Number of transactions of a given type paid for by respondents in the 24-hour period before the survey.
Bill Size	Amount of non-tip expenditure on a bill (e.g., sum of relevant menu prices).
Cash Bill	Yes if non-tipped expenditure was paid in cash, 0 otherwise.
Tipped Expenditure	Expenditure for a given transaction that takes the form of a tip.
Total Tipped Expenditure	Total Tipped Expenditure across all transactions of a given type.
Stiff Rate	Percentage of transactions in which there was no tipped expenditure.
Tip Rate	Ratio of Tipped Expenditure on Transaction over Non-Tipped Expenditure for a given Transaction.
Tipped Percentage	The ratio of the total tipped expenditure across all transactions of a given type to the total Bill Size across all transactions of that type.
Cash Tip	Yes if tipped expenditure was paid in cash, 0 otherwise.

[14] Chetty, R., Hendren, N., Kline, P., & Saez, E. (2014). Where is the land of opportunity? The geography of intergenerational mobility in the United States. *The Quarterly Journal of Economics*, *129*(4), 1,553–1,623.

National-Level Estimates

This section presents summary statistics concerning transactions in seven commonly tipped industries. Estimates of the frequency of transactions, tipped and non-tipped, in these industries are presented in Table 5. Partial and full-service restaurant transactions are the most frequent types of transactions in tipping-related industries, whereas home maintenance, casino, and hotel transactions are the least frequent. Note, however, that these less frequent types of transactions also have higher average bill sizes, and thus may be of greater economic importance than the number of transactions would indicate. The fraction of transactions for which the bill is at least partly paid in cash is lowest for full-service restaurants and generally higher in other industries. At the same time, the stiff rate, or the percentage of transactions without a tip, is by far the lowest for full-service restaurants, and the ratio of tipping to non-tipped (i.e., bill) expenditure, or the tip percentage, is also relatively high for full-service restaurants.

Table 5: Transaction Frequency and Characteristics by Industry (Standard Errors)

Industry	Mean # of Daily Transactions	Mean Bill Size	% of Bills Paid in Cash	Stiff Rate	Tip Percentage**
Full-Service Restaurants	0.16 (<0.01)	$46.86 ($1.63)	34% (1%)	5% (1%)	20% (<1%)
Partial-Service Restaurants	0.38 (<0.01)	$16.97 ($0.37)	46% (1%)	70% (1%)	8% (<1%)
Hotel*	0.02 (<0.01)	$104.20 ($11.32)	45% (3%)	41% (4%)	12% (3%)
Stylists/ Grooming	0.08 (<0.01)	$38.05 ($1.36)	55% (2%)	21% (1%)	18% (1%)
Home Maintenance	0.03 (<0.01)	$237.04 ($30.58)	51% (3%)	72% (2%)	4% (1%)
Casino*	0.02 (<0.01)	$85.69 ($25.56)	65% (3%)	25% (3%)	28% (11%)
Transportation	0.04 (<0.01)	$24.15 ($2.22)	44% (3%)	48% (2%)	17% (2%)

*Hotel and casino transactions are restricted to those with an associated bill (e.g., no valets or bellhops) for comparability with other transaction categories.

**Tip Percentage is an estimate of the ratio of total tipping expenditure to total non-tipped expenditure (i.e., bill excluding tips). It is the mean of the ratio of the total tips (including zero tips) over bill size across all transactions, in which the transactions are weighted by the product of poststratification weight and bill size.

Table 6 presents more detailed characteristics of tips by major industry. The relatively low probability that a full-service restaurant bill is paid in cash is also reflected in a relatively low probability that a tip in that industry is paid in cash. Full- and partial-service restaurants have the lowest mean tip rate compared to less frequent transaction types. When breaking the mean tip rate down by cash versus non-cash tips, there is little evidence that the tip rates are significantly different between the types of tips.

Tip rates for cash tips are generally higher than non-cash tips for full-service restaurants, however, although the estimated mean tip rates differ by less than two percentage points.

Table 6: Tip Characteristics by Industry (Standard Errors)

Industry	% of Tips Paid in Cash	Mean Tip Rate, All Tips	Mean Tip Rate, Non-Cash Tips	Mean Tip Rate, Cash Tips	Ho: Cash—Non-Cash =0 *T*-stat
Full-Service Restaurants	57% (1%)	0.215 (0.002)	0.207 (0.003)	0.222 (0.003)	3.51*
Partial-Service Restaurants	70% (1%)	0.196 (0.003)	0.194 (0.006)	0.199 (0.004)	0.74
Hotel	79% (3%)	0.462 (0.047)	0.573 (0.165)	0.420 (0.053)	-0.84
Stylists/ Grooming	75% (2%)	0.256 (0.006)	0.261 (0.012)	0.255 (0.006)	-0.44
Home Maintenance	83% (4%)	0.369 (0.041)	0.280 (0.087)	0.376 (0.053)	0.94
Casino	85% (3%)	0.385 (0.028)	0.412 (0.070)	0.354 (0.029)	-0.79
Transportation	80% (2%)	0.353 (0.017)	0.314 (0.035)	0.347 (0.024)	0.84

*Statistically significant at the 5% level.

To ascertain which industries are relatively important with respect to the generation of tipping income, Table 7 presents estimates of annual total tipped expenditure (in billions of $). Full-service restaurants received the largest amount of tipped expenditure, followed by partial-service restaurants and stylists/ grooming. The transportation industry received the least. Cash tipping appears to be especially important in the home maintenance and the casino industries.

Table 7: Total Annual Tipped Expenditure by Industry (Standard Errors)

Industry	Total Tipped Expenditure (Billions)	Total Tipped Expenditure (Billions), Cash Tips**	Total Tipped Expenditure (Billions), Non-Cash**
National	$334.43 ($25.21)	$207.55 ($20.32)	$94.95 (4.11)
Full-Service Restaurants	$132.78 ($4.84)	$71.90 ($3.35)	$55.46 ($2.07)
Partial-Service Restaurants	$47.61 ($2.32)	$31.28 ($1.47)	$13.65 ($0.54)
Hotel*	$19.50 ($3.32)	$8.09 ($1.05)	$5.87 ($1.42)
Stylists/ Grooming	$49.08 ($3.41)	$31.35 ($1.27)	$13.06 ($1.10)
Home Maintenance	$25.86 ($4.75)	$19.53 ($4.24)	$1.96 ($0.57)
Casino*	$45.12 ($16.80)	$36.88 ($16.58)	$2.81 ($1.10)
Transportation	$14.49 ($1.95)	$8.50 ($0.63)	$2.14 ($0.37)

*Hotel and casino transactions are restricted to those with an associated bill (e.g., no valets or bellhops) for comparability with other transaction categories.
**The sum of cash and non-cash expenditure will not necessarily sum to total expenditure because of nonresponse to the tip mode question.

The main results transactions are grouped into seven larger categories for the purpose of maintaining sample size and thus, statistical precision. In Appendix B, the stiff and tip rates are disaggregated in to types of transactions as presented in Table B1.

Commuting Zone Estimates

MRP estimates for national tipping rates, stiffing rates, and tipping percentage are presented in Table 8. These estimates are largely consistent with those estimates generated using the poststratification weights. This in turn indicates that the national estimates are robust to differences in poststratification methodology.

Table 8: MRP Estimates of National Tipping Rates, Stiffing Rates, and Tipping Percentage by Industry.

Industry	Tipping Rate		Stiffing Rate		Tipping Percent age	
	National Estimate	# of CZs Statistically Significantly Different from National Average	National Estimate	# of CZs Statistically Significantly Below National Average	National Estimate	# of CZs Statistically Significantly Below National Average
Full-Service Restaurants	0.215 (0.002)	0	0.047 (0.005)	0	0.204 (0.003)	0
Partial-Service Restaurants	0.196 (0.003)	1	0.704 (0.006)	12	0.088 (0.003)	1
Stylists/ Grooming	0.256 (0.005)	0	0.207 (0.013)	2	0.203 (0.007)	0
Transportation	0.366 (0.023)	0	0.485 (0.026)	2	0.102 (0.072)	0

Commuting zone estimates of the three tipping outcomes for each industry are presented in a separate spreadsheet. For the purpose of interpreting geographic patterns of tipping behavior, the focus of this section will be on the tipping percentage, which is a summary measure of the importance of tipping in a given industry, reflecting estimates of both the tipping rate and stiffing rate, and can be interpreted as the ratio of industry income from tipping to income from non-tipping (i.e. bills). Different industries display different geographic patterns with respect to the tipping percentage. Although commuting zones with the highest full-service restaurant tipping percentage are concentrated in the South east, tipping is a more important source of income for partial-service restaurants in the Northeast. Tipping revenue is predicted to be a more important source of revenue for grooming establishments in the Midwest than in other regions, whereas tipping is especially important to transportation services in the West.

However, it is important not to over-interpret differences in tipping behavior between commuting zones. The model parameters, and thus predictions, are subject to sampling variability. To generate individual tests for whether a commuting zone estimate is different from the national average, a test statistic is calculated as the ratio of the difference between the commuting zone estimate and the national estimate as well as the bootstrapped standard error of this estimated difference. A test statistic with an absolute value above 4 is taken as evidence of a statistically significant difference between the commuting zone estimate and the national estimate.[15] As can be seen in Table 6, only a handful of the 709 commuting zones have an estimate that is statistically significantly different than the national estimate, and then mostly with respect to the stiffing rate. Although there may still be geographic differences, under the assumed regression models, there is not enough information to support strong conclusions concerning differences between commuting zones.

Conclusion

This report documents results from the first six months of the IRS survey on consumer tipping behavior. Estimates are generated by industry both for the country as a whole, as well as by commuting zones. The results in Tables 5 and 6 are consistent with the existence of variability in tipping and stiffing rates between industries. Geographic variation in tipping and stiffing rates is subject to greater uncertainty due to sampling variability.

[15] The choice of approximately 4 as the threshold is the inverse normal of .025 divided by 709 to account for multiple comparisons.

Appendix A: Demographic and Geographic Characteristics

Table A1: Age by Gender by Region by Income Balancing

Sampling Cell	Balancing %
Male 18–34 Northeast Under $20K	0.333
Male 18–34 Northeast $20K–$49.9K	0.675
Male 18–34 Northeast $50K–$99.9K	1.095
Male 18–34 Northeast $100K+	0.605
Male 18–34 Midwest Under $20K	0.442
Male 18–34 Midwest $20K–$49.9K	0.898
Male 18–34 Midwest $50K–$99.9K	1.219
Male 18–34 Midwest $100K+	0.562
Male 18–34 South Under $20K	0.79
Male 18–34 South $20K–$49.9K	1.604
Male 18–34 South $50K–$99.9K	1.985
Male 18–34 South $100K+	1.004
Male 18–34 West Under $20K	0.491
Male 18–34 West $20K–$49.9K	1.043
Male 18–34 West $50K–$99.9K	1.33
Male 18–34 West $100K+	0.655
Male 35–54 Northeast Under $20K	0.286
Male 35–54 Northeast $20K–$49.9K	0.825
Male 35–54 Northeast $50K–$99.9K	1.351
Male 35–54 Northeast $100K+	1.07
Male 35–54 Midwest Under $20K	0.34
Male 35–54 Midwest $20K–$49.9K	0.955
Male 35–54 Midwest $50K–$99.9K	1.486
Male 35–54 Midwest $100K+	0.863
Male 35–54 South Under $20K	0.646
Male 35–54 South $20K–$49.9K	1.641
Male 35–54 South $50K–$99.9K	2.453
Male 35–54 South $100K+	1.463
Male 35–54 West Under $20K	0.396
Male 35–54 West $20K–$49.9K	0.992
Male 35–54 West $50K–$99.9K	1.531
Male 35–54 West $100K+	1.011
Male 55+ Northeast Under $20K	0.351
Male 55+ Northeast $20K –$49.9K	1.091
Male 55+ Northeast $50K –$99.9K	0.991
Male 55+ Northeast $100K+	0.572
Male 55+ Midwest Under $20K	0.391

Sampling Cell	Balancing %
Male 55+ Midwest $20K –$49.9K	1.326
Male 55+ Midwest $50K –$99.9K	1.264
Male 55+ Midwest $100K+	0.608
Male 55+ South Under $20K	0.75
Male 55+ South $20K –$49.9K	2.166
Male 55+ South $50K –$99.9K	2.04
Male 55+ South $100K+	1.066
Male 55+ West Under $20K	0.448
Male 55+ West $20K –$49.9K	1.183
Male 55+ West $50K –$99.9K	1.174
Male 55+ West $100K+	0.626
Female 18–34 Northeast Under $20K	0.311
Female 18–34 Northeast $20K–$49.9K	0.656
Female 18–34 Northeast $50K–$99.9K	1.001
Female 18–34 Northeast $100K+	0.516
Female 18–34 Midwest Under $20K	0.415
Female 18–34 Midwest $20K–$49.9K	0.846
Female 18–34 Midwest $50K–$99.9K	1.335
Female 18–34 Midwest $100K+	0.565
Female 18–34 South Under $20K	0.745
Female 18–34 South $20K–$49.9K	1.5
Female 18–34 South $50K–$99.9K	2.352
Female 18–34 South $100K+	1.095
Female 18–34 West Under $20K	0.474
Female 18–34 West $20K–$49.9K	1.021
Female 18–34 West $50K–$99.9K	1.413
Female 18–34 West $100K+	0.662
Female 35–54 Northeast Under $20K	0.24
Female 35–54 Northeast $20K–$49.9K	0.784
Female 35–54 Northeast $50K–$99.9K	1.209
Female 35–54 Northeast $100K+	0.696
Female 35–54 Midwest Under $20K	0.295
Female 35–54 Midwest $20K–$49.9K	0.934
Female 35–54 Midwest $50K–$99.9K	1.714
Female 35–54 Midwest $100K+	0.915
Female 35–54 South Under $20K	0.555
Female 35–54 South $20K–$49.9K	1.625
Female 35–54 South $50K–$99.9K	3.054
Female 35–54 South $100K+	1.851

Sampling Cell	Balancing %
Female 35–54 West Under $20K	0.357
Female 35–54 West $20K–$49.9K	0.982
Female 35–54 West $50K–$99.9K	1.745
Female 35–54 West $100K+	1.162
Female 55+ Northeast Under $20K	0.389
Female 55+ Northeast $20K –$49.9K	1.385
Female 55+ Northeast $50K–$99.9K	1.203
Female 55+ Northeast $100K+	0.579
Female 55+ Midwest Under $20K	0.484
Female 55+ Midwest $20K –$49.9K	1.642
Female 55+ Midwest $50K –$99.9K	1.486
Female 55+ Midwest $100K+	0.632
Female 55+ South Under $20K	0.844
Female 55+ South $20K–$49.9K	2.683
Female 55+ South $50K –$99.9K	2.413
Female 55+ South $100K+	1.125
Female 55+ West Under $20K	0.462
Female 55+ West $20K –$49.9K	1.518
Female 55+ West $50K –$99.9K	1.373
Female 55+ West $100K+	0.672

Table A2: Poststratification Variables, Weighted and Unweighted Proportions (Sources for Target Proportions, N = 18,032)

Category	Unweighted Proportion	Weighted Proportion
Rural–Urban Continuum Codes (ACS 5-Year)		
Counties in metro areas of 1 million population or more	54.08%	54.90%
Counties in metro areas of 250,000 to 1 million population	22.23%	21.17%
Counties in metro areas of fewer than 250,000 population	9.64%	9.18%
Non-metro counties	14.06%	14.75%
Census Region (ACS 5-Year)		
Northeast	21.64%	18.09%
Midwest	24.48%	21.31%
South	34.22%	37.30%
West	19.65%	23.29%
Educational Attainment (ACS 5-Year)		
High School or Less	22.24%	41.50%
Some College	37.25%	31.30%
College	26.14%	17.31%
Graduate Degree	14.37%	9.89%
Gender (ACS 5-Year)		

Category	Unweighted Proportion	Weighted Proportion
Female	57.73%	51.37%
Male	42.27%	48.63%
Age (ACS 5-Year)		
18–24	6.79%	12.92%
25–34	15.86%	17.66%
35–44	13.83%	16.74%
45–64	43.06%	34.31%
65+	20.47%	18.37%
Race/Ethnicity (Ipsos Poststratification Weights)		
Non-Hispanic White	80.91%	69.43%
Non-Hispanic Black	7.45%	12.57%
Hispanic	5.77%	9.87%
Other Non-Hispanic	5.87%	8.13%
Day of the Week of Survey Start Date		
Sunday	11.57%	14.29%
Monday	14.40%	14.29%
Tuesday	13.91%	14.29%
Wednesday	14.74%	14.29%
Thursday	15.63%	14.29%
Friday	16.85%	14.29%
Saturday	12.92%	14.29%
Month of Survey Start Date		
January	21.35%	17.13%
February	16.54%	15.47%
March	16.54%	17.13%
April	15.39%	16.57%
May	15.58%	17.13%
June	14.60%	16.57%

Table A3: MRP Poststratification Variables, Population Proportions (Sources for Target Proportions , *N* = 18,028)

Category	Sample Proportion	Population Proportion
Individual-Level Variables		
Gender (Reference: Female)		
Male	42.27%	48.63%
Age (Reference: 18-24)		
25–34	15.86%	17.66%
35–44	13.83%	16.74%
45–64	43.06%	34.31%
65+	20.47%	18.37%
Educational Attainment (Reference: High School or Less)		
Some College	37.25%	31.30%
College	26.14%	17.31%
Graduate Degree	14.37%	9.89%
County-Level Variables		
% of County Foreign Born	14.49%	16.13%
Race/Ethnic Composition (Reference: % of County Population, White)		
% of County Population , Black	11.63%	12.25%

Category	Sample Proportion	Population Proportion
% of County Population , Hispanic	14.74%	16.89%
% of County Population, Other	7.57%	8.29%
% of County Households by Income Bracket (Reference: % of Households <$10,000)		
% of Households, $10,000 – $14,999	5.21%	5.25%
% of Households, $15,000 – $24,999	10.57%	10.54%
% of Households, $25,000 – $34,999	10.14%	10.05%
% of Households, $35,000 – $49,999	13.47%	13.36%
% of Households, $50,000 – $74,999	17.92%	17.78%
% of Households, $75,000 – $99,999	12.18%	12.12%
% of Households, $100,000 – $149,999	13.15%	13.20%
% of Households, $150,000+	10.25%	10.51%
Rural–Urban Continuum Codes (Reference: Counties in metro areas of 1 million population or more)		
Counties in metro areas of 250,000 to 1 million population	22.23%	21.17%
Counties in metro areas of fewer than 250,000 population	9.64%	9.18%
Non-metro counties	14.06%	14.75%
Census Region (Reference: Northeast)		
Midwest	24.48%	21.31%
South	34.22%	37.30%
West	19.65%	23.29%

Appendix B : National Tipping Outcomes

Table B1: Stiff Rate and Tip Rates by Sub -Industry

Sub-Industry	Stiff Rate			Tip Rate		
	N	Estimate	Standard Error	*N*	Estimate	Standard Error
Full-Service Restaurant Transactions						
1: Full-Service Dining	3,079	5%	1%	2,933	0.215	0.002
Partial-Service Restaurants						
2: Fast Casual	1,410	53%	2%	646	0.203	0.006
3: Fast Food	3,256	86%	1%	388	0.193	0.006
4: Carryout/ Delivery	1,032	53%	2%	475	0.176	0.006
5: Bar	147	11%	3%	133	0.224	0.011
6: Coffee Shops	547	65%	3%	191	0.222	0.010
7: Ice Cream/ Smoothie Shops	79	65%	7%	23	0.187	0.032
8: Self-Service/ Cafeteria/ Buffets	180	46%	5%	88	0.186	0.013
9: Food Car/ Truck	26	53%	12%	11	0.202	0.019
Hotel Transactions						
10: Concierge/ Front Desk Staff	185	84%	4%	12	0.422	0.136
11: Housekeeping	229	58%	4%	33	0.641	0.128
12: Room Service	137	44%	5%	44	0.437	0.094
13: Valet	29	29%	10%	15	0.373	0.091
14: Bellhop/ Luggage Assistance	18	12%	7%	5	0.384	0.157
15: Bar	39	48%	11%	14	0.336	0.052
16: Full-Service Dining	92	35%	6%	48	0.457	0.083
17: Self-Service/ Cafeteria/ Buffets	76	66%	8%	8	0.304	0.129
18: Shuttle Service to/ from Hotel/ Motel	33	49%	10%	4	0.753	0.288
Stylists/ Grooming Transactions						
19: Hair Stylist	689	16%	2%	594	0.251	0.007
20: Barber	332	22%	3%	272	0.288	0.014
21: Manicurist/ Pedicurist	217	20%	4%	183	0.209	0.013
22: Massage Therapist	94	17%	5%	75	0.231	0.017
23: Waxing/ Hair Removal	56	21%	8%	49	0.301	0.033
24: Facial/ Skin Care	69	66%	7%	26	0.242	0.044
25: Makeup Artist	6	80%	19%	1		
Home Maintenance Transactions						
26: Professional Movers	46	33%	7%	30	0.391	0.094
27: Maid or Cleaning Service	100	53%	6%	41	0.541	0.097
28: Lawn/ Gardening Service	151	75%	5%	34	0.278	0.030
29: Handyman/ Repairman	169	85%	3%	27	0.144	0.032
30: Equipment Rental	46	91%	5%	5	0.759	0.152

Sub-Industry	Stiff Rate			Tip Rate		
	N	Estimate	Standard Error	*N*	Estimate	Standard Error
31: Dealers	140	40%	5%	0		
32: Floor Servers	157	39%	5%	44	0.561	0.068
Casino Transactions						
33: Bar	155	22%	4%	91	0.379	0.060
34: Full-Service Dining	98	27%	5%	54	0.376	0.057
35: Self-Service/Cafeteria/Buffets	118	47%	5%	51	0.258	0.036
36: Shuttle Service to/from Casino	24	40%	11%	4	0.372	0.075
37: Valet	39	23%	8%	14	0.413	0.090
Transportation Transactions						
38: Limousine	22	23%	8%	16	0.488	0.087
39: Standard Taxi (e.g., "yellow cabs")	221	37%	4%	151	0.293	0.017
40: Uber, Lyft, or other Ride-Share Service	358	55%	3%	165	0.391	0.025
41: Shuttle Service	41	50%	10%	19	0.297	0.080
42: Valet	8	40%	19%	5	0.627	0.191

Appendix C: National Tipping Outcomes —Excluding Only Full -Service Restaurant Outliers

Table C1: Transaction Frequency and Characteristics by Industry (Standard Errors)

Industry	Mean # of Daily Transactions	Mean Bill Size	% of Bills Paid in Cash	Stiff Rate	Tip Percentage**
Full-Service Restaurants	0.16 (<0.01)	$48.91 ($1.79)	34% (1%)	4% (<1%)	20% (<1%)
Partial-Service Restaurants	0.55 (0.01)	$40.84 ($13.10)	47% (1%)	63% (1%)	205% (139%)
Hotel*	0.04 (<0.01)	$271.64 ($157.89)	45% (3%)	28% (3%)	82% (12%)
Stylists/ Grooming	0.14 (0.01)	$165.17 ($90.20)	56% (1%)	19% (1%)	178% (126%)
Home Maintenance	0.06 (<0.01)	$261.95 ($33.25)	52% (2%)	50% (3%)	39% (10%)
Casino*	0.04 (<0.01)	$457.89 ($335.88)	54% (3%)	20% (2%)	31% (3%)
Transportation	0.07 (0.01)	$48.47 ($7.72)	44% (2%)	39% (3%)	72% (15%)

*Hotel and casino transactions are restricted to those with an associated bill (e.g., no valets or bellhops) for comparability with other transaction categories.

**Tip Percentage is an estimate of the ratio of total tipping expenditure to total non-tipped expenditure (i.e., bill excluding tips). It is the mean of the ratio of the total tips (including zero tips) over bill size across all transactions, in which the transactions are weighted by the product of poststratification weight and bill size.

Table C2: Tip Characteristics by Industry (Standard Errors)

Industry	% of Tips Paid in Cash	Mean Tip Rate, All Tips	Mean Tip Rate, Non-Cash Tips	Mean Tip Rate, Cash Tips	Ho: Cash—Non-Cash =0 *T*-stat
Full-Service Restaurants	59% (1%)	0.214 (0.002)	0.206 (0.003)	0.221 (0.003)	3.76*
Partial-Service Restaurants	70% (1%)	5.106 (4.574)	16.753 (16.151)	0.502 (0.068)	-1.01
Hotel	69% (5%)	1.053 (0.099)	1.601 (0.228)	0.835 (0.068)	-3.49*
Stylists/ Grooming	75% (2%)	0.989 (0.380)	2.323 (1.589)	0.541 (0.051)	-1.13
Home Maintenance	64% (4%)	1.386 (0.129)	2.092 (0.206)	1.124 (0.166)	-4.18*
Casino	70% (4%)	1.737 (0.814)	1.603 (0.276)	1.844 (0.938)	0.25
Transportation	65% (3%)	1.555 (0.515)	3.642 (1.698)	0.668 (0.064)	-1.75

*Statistically significant at the 5% level.

Table C3: Total Annual Tipped Expenditure by Industry (Standard Errors)

Industry	Total Tipped Expenditure (Billions)	Total Tipped Expenditure, Cash Tips (Billions)**	Total Tipped Expenditure, Non-Cash (Billions)**
National	$9,610.05	$3,088.07	$6,296.51
	($3619.72)	($1,315.23)	($3,228.49)
Full-Service Restaurants	$137.13	$75.53	$55.54
	($5.27)	($3.29)	($2.58)
Partial-Service Restaurants	$3,986.23	$1,332.59	$2,591.84
	($2,506.62)	($855.84)	($2,341.68)
Hotel*	$823.79	$137.25	$656.26
	($595.34)	($52.05)	($573.23)
Stylists/ Grooming	$3,454.82	$866.47	$2,550.30
	($2,331.08)	($677.91)	($2,273.53)
Home Maintenance	$491.31	$230.85	$219.47
	($117.53)	($83.25)	($73.39)
Casino*	$501.26	$390.79	$85.02
	($328.38)	($307.12)	($21.15)
Transportation	$213.69	$51.94	$135.93
	($63.78)	($14.44)	($48.43)

*Hotel and casino transactions are restricted to those with an associated bill (e.g., no valets or bellhops) for comparability with other transaction categories.

** The sum of cash and non-cash expenditure will not necessarily sum to total expenditure because of non-response to the tip mode question.

Table C4: Stiff Rate and Tip Rates by Sub-Industry

Sub-Industry	Stiff Rate			Tip Rate		
	N	Estimate	Standard Error	*N*	Estimate	Standard Error
Full-Service Restaurant Transactions						
1: Full-Service Dining	3,487	4%	<1%	3,319	0.214	0.002
Partial-Service Restaurants						
2: Fast Casual	2,305	48%	1%	1,153	15.802	15.455
3: Fast Food	4,845	80%	1%	872	1.380	0.657
4: Carryout/ Delivery	1,692	48%	1%	848	0.477	0.150
5: Bar	300	9%	2%	274	0.743	0.234
6: Coffee Shops	1,078	58%	2%	412	0.550	0.076
7: Ice Cream/ Smoothie Shops	259	52%	4%	109	0.510	0.072
8: Self-Service/ Cafeteria/ Buffets	306	42%	4%	158	0.287	0.027
9: Food Car/ Truck	82	44%	6%	43	0.452	0.110
Hotel Transactions						
10: Concierge/ Front Desk Staff	320	72%	5%	61	1.172	0.171
11: Housekeeping	390	53%	3%	84	0.856	0.094
12: Room Service	325	30%	3%	157	1.055	0.132
13: Valet	84	23%	4%	50	1.599	0.598
14: Bellhop/ Luggage	52	6%	3%	30	1.280	0.204

Sub-Industry	Stiff Rate			Tip Rate		
	N	Estimate	Standard Error	*N*	Estimate	Standard Error
Assistance						
15: Bar	107	31%	6%	59	1.103	0.154
16: Full-Service Dining	172	29%	5%	95	0.787	0.108
17: Self-Service/ Cafeteria/ Buffets	135	60%	5%	29	0.873	0.211
18: Shuttle Service to/ from Hotel/ Motel	53	45%	8%	12	1.064	0.236
Stylists/ Grooming Transactions						
19: Hair Stylist	1,136	15%	1%	979	1.415	0.923
20: Barber	556	21%	3%	456	0.620	0.070
21: Manicurist/ Pedicurist	458	16%	3%	396	0.464	0.063
22: Massage Therapist	194	14%	2%	162	0.855	0.156
23: Waxing/ Hair Removal	156	18%	4%	133	1.192	0.349
24: Facial/ Skin Care	151	42%	6%	80	0.878	0.164
25: Makeup Artist	21	14%	9%	15	1.563	0.451
Home Maintenance Transactions						
26: Professional Movers	194	20%	4%	151	1.598	0.226
27: Maid or Cleaning Service	291	29%	4%	196	1.686	0.271
28: Lawn/ Gardening Service	274	58%	4%	106	1.051	0.145
29: Handyman/ Repairman	293	73%	3%	78	0.784	0.128
30: Equipment Rental	90	74%	8%	26	1.421	0.240
31: Dealers	208	38%	4%	0		
32: Floor Servers	270	30%	4%	113	1.037	0.184
Casino Transactions						
33: Bar	315	21%	3%	197	0.811	0.094
34: Full-Service Dining	224	22%	3%	139	0.981	0.220
35: Self-Service/ Cafeteria/ Buffets	197	36%	3%	103	0.643	0.105
36: Shuttle Service to/ from Casino	53	26%	5%	22	20.438	16.283
37: Valet	65	18%	6%	30	2.436	1.284
Transportation Transactions						
38: Limousine	93	15%	4%	76	1.484	0.134
39: Standard Taxi (e.g., "yellow cabs")	461	29%	3%	335	0.791	0.080
40: Uber, Lyft, or other Ride-Share Service	677	48%	3%	363	2.304	1.125
41: Shuttle Service	92	38%	7%	60	1.098	0.243
42: Valet	17	26%	16%	13	1.443	0.338

Appendix D: Commuting Zone Estimates

Please see attached Excel spreadsheet for commuting zone estimates.

Survey of Consumer Tipping Behavior: Technical Report

Prepared for the Internal Revenue Service

Prepared by Fors Marsh Group, LLC

March 2018

Version 2

Table of Contents

Introduction

The Internal Revenue Service (IRS) is charged with enforcing federal tax laws in the United States, including the lawful collection of tax revenue. This mission is complicated by economic activity that involves substantial cash-based transactions or other income that is not independently reported to the IRS and is difficult for the IRS to detect. Tipping that occurs in service industries is one such form of economic activity that poses a challenge for tax administrators.

To help address this challenge, IRS has undertaken a multi-year project culminating in a one-year survey of consumer tipping behavior in order to update and enhance its understanding of taxpayer compliance with respect to tip income reporting. Information on tipping behavior can be used by IRS to produce estimates of aggregate tip income across industries and regions, which can in turn be used to estimate unreported tip income.

Earlier phases of the multi-year project covered the development of the survey questionnaire, the survey methodology, and the determination of the sampling frame. An interim report with national and regional industry estimates of tipping behavior and tipping income based on the first six months of data collection has also been produced. This report focuses on the survey administration, data cleaning and the development of poststratification weights and accompanies the final, 12-month sample of survey data delivered to the IRS.

Study Background

In the first phase of the multi-year project, Fors Marsh Group (FMG) conducted a review of the literature on tipping behavior, identified options for key study elements such as sampling source, sampling mode, study design, and data analysis, and developed recommendations for proceeding. The recommendations included conducting a survey with a repeated cross-sectional design using an internet-based panel sample and survey questionnaire. The report that was produced from this phase of work, *Estimating Consumer Tipping Behavior: Review and Recommendations* (February 2014), is embedded in Appendix A.

Next, IRS worked with Fors Marsh Group to design the questionnaire for surveying consumers about their tipping behavior. Draft versions of the survey underwent cognitive testing and usability testing in which all elements of the instrument were rigorously examined, including the final list of industries or services to include, how the survey should be organized and worded for minimum confusion, cross-platform compatibility, and recall length. The results of the testing led to refinements of the survey questionnaire, which was then piloted in the next phase of the project. The report that describes the development of the survey in detail, *IRS Tipping Report on Cognitive and Usability Testing* (January 2015) is embedded in Appendix B.

In the next phase of work, FMG conducted a one-month pilot of the survey questionnaire with the primary goal of arbitrating between two sampling strategies – probability or non-probability. 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. Although the study showed no sample being more biased than the other, the results supported the use of the non-probability sample. Specifically, given the 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 and 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 vendor providing the non-probability sample in the pilot was Ipsos, which was selected for use in the yearlong survey described in this report. 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* (November 2015), is embedded in Appendix D.

Finally, Fors Marsh Group prepared a report using the first six months of data collected by the consumer tipping survey. The report presents national estimates of total tip expenditures, stiffing rates and tipping percentages by industry/occupation, method of payment, geographic region, and certain other factors. The report outlines FMG's recommended methodology for producing such estimates and provides a useful blueprint for developing similar estimates based on the full year of data. Appendix E contains an embedded version of this report, titled *Interim Report on the Survey of Consumer Tipping Behavior* (January 2018).

Survey Questionnaire and Administration

The final version of the screener and survey that was administered for the yearlong data collection is shown in Appendix C. The original versions of the survey, as well as the outcomes of cognitive testing done on the survey can be found in Appendix B.

Survey invitations, which included invitation text, a link to the survey program, and a link to the panel provider's member policies (including confidentiality), followed the standard email invitation formats used by Ipsos and each of its partners so that sampled individuals were familiar with how to use them to access the survey. Email invitations were sent out to registered panelists daily during the yearlong fielding to collect information about the prior day's transactions. Because there were no client quotas, the standard Ipsos demographic screener was used.

The survey instrument, estimated to take approximately 8 minutes or less, asked if respondents had transactions in the last calendar day such as:

- Restaurant or other prepared food/drink service
- Hotel/motel
- Personal grooming, beauty, or massage services
- Moving or household maintenance services
- Casino
- Taxi, limousine, rideshare, or shuttle service

For each category of transactions follow up questions were asked to specify what type of service was received at a more granular level. For each type of service, participants were asked what type of payment was used (cash, debit, credit, etc.), the amount of the bill excluding tip, if the business added an automatic tip, and if the participant left a voluntary tip for the transaction. The items further ask about the payment method and amount of the tip given.

Sample Design: Yearlong Survey

Target Population

The target population for the full, yearlong IRS tipping study was the same as for the one month pilot. It included all U.S. residents who used services that are commonly tipped. The number of individuals in this population is unknown but likely includes a majority of the U.S. adult population. Example settings where tipping is typical include full-service restaurants, taxis, barbershops, beauty salons, hotels, and casinos.

Sample Size

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, were 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 1,200 tipping occurrences per industry over the course of a year or to average 100 tipping occurrences per month for each industry. Table 1 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 the 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 were higher than indicated in other sources, and thus, resulted in more tipping incidences for a given sample. However, given that the pilot study was conducted for one summer month, the incidence rates in the table might not be representative of what would be obtained from a yearlong fielding period. Consequently, to be conservative, the incidence rates in the table should be interpreted as upper bounds, 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 was "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 was 57,143. This determined our target sample size for the full yearlong survey, which was 60,000 completed responses.

Table 1: 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

Non-probability Sample

For the 2017 survey administration, the non-probability sample was collected from Ipsos' blended panel, which was also used in the one-month pilot previously administered. Details of the Ipsos sample are presented in the following section.

Sampling Methodology

Ipsos' blended sample approach combined the use of its Ampario online sampling method with its i-Say online panel—an online panel of 800,000 members and their households. Ampario is a non-probability sampling procedure developed by Ipsos that invites respondents by invitations, banner ads, and other means on 100 to 400 websites that have partnered with Ipsos. These two methods were combined into a single sample using Ipsos' proprietary Cortex routing system, which allocates and reallocates a sample based on respondent eligibility. Simply put, when respondents were not eligible for one survey, they were immediately redirected to other surveys in progress. In traditional one-off, opt-in surveys, ineligible respondents are lost, representing a considerable cost. Finally, Bayesian methodology, which requires previous information regarding the overall sample of interest in order to mix with current information for the final distribution of results, was used to form the final distribution. As is the case with a traditional online sample, Ipsos' blended sampling can work with several different data collection modes, but it is best served with an online-based questionnaire, which included a longitudinal diary approach for this study. However, because of the opt-in nature of the blended sample, it is not possible to model the probability of responding to a survey, thus there exists a source of potential bias in survey estimates.

Recruitment Sources Used in the Project

Ipsos i-Say	Ipsos' panels are not just lists or databases of individuals, but they are actively managed research Access Panels: • Individuals who have volunteered to take part in market research surveys • Created and managed for long-term use and access • Extensively profiled to efficiently target respondents The vast majority of panelists are referred to Ipsos through various online suppliers. Ipsos only uses high-quality recruitment sources to entice people who are eager to take surveys. The organization strategically focuses on developing processes that reflect the newest internet practices, as may currently be found through social networks. Email lists, banners, website and text ads, co-registration, and search engine marketing are also used.
Lightspeed	This is an actively managed panel composed of people who made a conscious

GMI	decision to participate in online surveys through a double opt-in registration process. Several methodologies are used to recruit panelists, including opt-in email, co-registration, e-newsletter campaigns, and traditional banner placements, as well as both internal and external affiliate networks. Social media is included through Lightspeed's recruiting partners.
Market Cube	Market Cube owns and operates the Univox Community, an actively managed panel with an individual-level compensation model. Market Cube also has access to a vast network of social media and publisher respondents that can be used to supplement internal assets. Additionally, Market Cube has developed close relationships with a variety of panel companies with which they can partner on difficult-to-reach subpopulations. These strategic partnerships allow Market Cube to leverage relevant lists, databases, and networks to fulfill specific client requirements.
ROIRocket	This large ad network has provided more than 30 million panelists to date and offers access to more than 5 million active respondents at any given time. The company has experience using its sample for online communities, custom panels, in-depth interviews, and longitudinal research studies.
SSI	SSI actively manages this panel and incorporates participants from partnership sources. Participants are recruited via banners, invitations, and messaging. Prospects go through rigorous quality controls before being included in SSI panels.

Quota Sampling Methods and Variables

Sample Balancing

Ipsos and each of its partners selected what is known as a "balanced return" sample, wherein the demographic distribution of "clicks" (meaning respondents who respond to a survey invitation by clicking the hyperlink and entering the survey) matches the demographic distribution of the overall U.S. population, as indicated in most recent results of the Census Bureau's Current Population Survey (CPS).[1] Because different individuals and demographic groups respond at different rates, the different sampling rates are applied for these different groups. The demographic distribution of the contacted sample, thus, does not match the demographic distribution of the U.S. population.

[1] To ensure sufficient sample records to complete the necessary number of interviews each month, multiple sample sources were needed. The sample for the IRS Consumer Tipping Study was provided by Ipsos' opt-in i-Say panel and four other opt-in panels, with the anticipated proportion of completed interviews provided by each source remaining constant each month (and following the proportions used in the pilot test). Each panel provider prepared responses to ESOMAR's 28 questions for online samples and was vetted by Ipsos' online research department. These panel providers emailed invitations to their panelists with a link that directed them to the Ipsos survey site after passing them through an intermediary site used by the panel provider to monitor whether panelists (a) responded and (b) completed the survey, so that their traditional panel incentive could be paid. Panel partners provided information on how many invitations were sent and balanced their samples using targets provided by Ipsos.

Ipsos undertook sample balancing (i.e., determining the proportion of the sample to allocate to different demographic groups), which was completed using four demographic variables: gender, age, region, and income. The links between each of these characteristics and tip rates have been the subject of past academic studies on tipping behavior. These variables were fully crossed, creating 96 sampling cells (see Table F1 in Appendix F). The levels (sample groups) within each of the variables are indicated in Table 2.

Table 2: Stratification Variables

Gender	Age	Region	Income
(1) Male	Age 18–34	(1) Northeast	Under $20K
(2) Female	Age 35–54	(2) Midwest	$20K–$49,999
	Age 55+	(3) South	$50K–$99,999
		(4) West	$100K+

Ipsos selected samples two times a week (Monday and Friday). On Monday, the sample was designed to produce a demographically balanced return sample equal to four days of completed interviews. On Friday, the sample was designed to produce the balanced return sample equal to three days of completed interviews. The samples were divided into replicates or subsamples that equally represented the larger sample (four replicates for the Monday samples; three replicates for the Friday samples), so that one replicate could be "released" (meaning survey invitations were sent to those sampled individuals) each day. This approach yielded approximately the targeted 144 daily completed interviews.

This approach of using sample replicates is employed to achieve greater efficiency when many sample balancing cells are employed by ensuring higher response rates in relatively sparse sampling cells.

The sample design assumed a one-month reuse of sample (i.e., individuals who were sampled for the study during one month were ineligible for another contact until the next month).

Ipsos employed a number of quality checks during the data collection process.

- Survey level:
  - Filtered respondents based on participation history
  - Screened respondents based on demographic variables being captured for the survey (age, gender, ZIP code, etc.)

- Engine level:
  - GeoIP verification: validated survey country versus respondent country determined by IP address
  - Language verification: validated survey language versus respondent language
  - Device check: matched between device used by respondent and the device setting of the survey
  - Used an algorithm to identify possibly unengaged respondents (straight-lining, speeding, providing invalid verbatim in open-ended questions)
  - Concurrent sniff-out session: filtered respondents with more than one opened session in the same browser on the same survey

- Fraud Profile Flag 4 (FPF4): determined mismatch using machine time versus time based on geolocation
- Open and anonymous proxy checks
- VOID: analyzed web cookies, PanelistID/SupplierID (identifiers provided by sample sources), RelevantID (third-party security service), SHA-1 hash function

Quality Assurance Procedures

The FMG team followed several quality assurance procedures both before the survey officially began and throughout the year to ensure that the survey was gathering data as intended.

Web Survey Quality Control. Ipsos programmed the survey based on upon the final survey questionnaire with programming notes and exclusions provided. Using the questionnaire, a quality assurance team reviewed the scripting of the survey to confirm correct programming. The programmed survey was then made available for review on a staging site for final reviewers at Ipsos and FMG. Before the survey began, the FMG team performed full testing of the programmed instrument to ensure that skip logic, randomization, conditional data piping, question wording, and all other specifications for the survey instrument were met. FMG's online survey quality control process was thorough and included checks to ensure that there were no grammatical or formatting errors, that the question type was accurate (single punch vs. multi-punch, etc.), that skip patterns functioned appropriately, and that data restrictions for open-ended questions matched requirements. The FMG team also had data capture checks in place to examine the functionality of the programmed survey. As a standard quality control check, multiple FMG researchers responded to the online survey and simultaneously recorded their answers on a paper copy of the survey; during these checks, researchers tested all branching/skip patterns in the questionnaire.

Data Collection. In fulfilling the contract, the FMG team was required to provide weekly updates on survey completions to the IRS. Ipsos gathered this data from their vendors and FMG conducted checks and ran trends to ensure that there were no anomalies or to identify any seasonal patterns occurring throughout survey fielding. This was how some vendor invitation issues had been identified early in the survey fielding (see Timeline of Survey Administration Issues section).

Data Quality Control. Ipsos cleaned any cases without key demographic variables that were needed for their weighting procedure such as region, race, and gender. Ipsos implemented procedures to identify fraudulent completes, and as a result, removed 20 cases that were deemed fraudulent. In this case, fraudulent cases were respondents who were considered to take the survey three or more times faster than the median speed per survey or respondents who provided the same response throughout at least one grid formatted question and completed the survey two times faster than the median respondent. Data cleaning steps beyond the initial cleaning are outlined later in the Data Cleaning section of this report.

Table 3: Survey Session by Status

Status	Definition	Frequency

Abandoned	Drop out – Respondent accessed the link but did not finish the survey .	339
Screen-out Client	The respondent was screened out based on survey screener questions.	26,050
Quota Full Client	Technological issue with Cortex	1
Complete	Valid respondent completed the survey .	18,099
Error	Respondents experienced technological errors with the link.	1
Wrong Complete	Completes that occurred due to vendor oversampling	363
Total		44,853

Data Verification and Cleaning. Once data collection was completed and all survey data entered, the data sets were reviewed and thoroughly checked before the final data were delivered to the IRS . Records were inspected to determine whether any completed cases should have been discarded. These data quality control checks were made to ensure that the analysis file was clean. Table 4 details the steps taken.

Table 4: Data Cleaning Steps

Data Cleaning Steps Taken Before Analysis	
1) Receive data sets	9) Check skip patterns
2) Print format library (file information)	10) Check recodes
3) Run frequencies (weighted and unweighted)	11) Check calculated variables
4) Check variable names	12) Check coding of “other, specify”
5) Check variable labels	13) Address problems
6) Check value labels	14) Make changes to formats
7) Check weights (against known pop. totals)	15) Secondary review of final data set
8) Check unweighted sampling	16) Recheck all resultant values

Timeline of Survey Administration Issues

The sampling issues that were encountered during fielding were largely a result of errors in vendor adherence to monthly quotas. As an error was discovered, Ipsos acted to identify the root of the issue and, ultimately, to provide a timely and fitting solution.

January Wave: The IRS Tipping Study was launched on January 2, 2017, and after the first week of fielding, it was discovered that the vendor Lightspeed (LSR) did not apply quotas when launching the sample. Lightspeed, in conjunction with Global Market Insite (GMI), was expected to achieve 10% of the 1,500 monthly completes, but due to the missing quotas and accelerated completion rates, the majority of their monthly completes were achieved in the first two days of field. The resulting completes were invalidated, and the vendor was requested to spread their full monthly quota across the remaining three weeks of the wave. Ipsos informed the vendor that their survey strategy needed to be amended and quotas properly applied. However, the same issue occurred the following week, as the quotas that LightSpeed applied did not hold. As a result, Ipsos removed 274 subsequent completes that were achieved through LSR/GMI for the first two weeks of fielding. LightSpeed was asked to spread their full monthly quota across the remaining two weeks of the January wave. Completes were monitored for the rest of the month to ensure that there were no further issues with sampling, and the wave was successfully closed, having achieved 1,500 completes. FMG informed IRS of these issues and requested that the 274 completes be labeled in the final data files (labeled as "Wrong Completes" in the STATUS variable).

April Wave: The vendor SSI/Opin. failed to achieve its monthly quota of completes. The anticipated proportion of completed surveys from SSI/Opin. should have been 25% (375 completes), and the actual achieved proportion was roughly 24%, resulting in a loss of 11 completes. Because the four other opt-in vendors attained their monthly quota with exact figures, the overall wave did not meet the monthly quota of 1,500 completes. This issue was due to a reduction in outgo, or outgoing invitations, during the wave. Previously, SSI/Opin. had been sampling at a much greater degree and dropped its sample outgo by roughly 40,000 invites. Because of the drastic decrease in outgo, the estimated number of completes based on the preceding waves was not achieved. Ipsos communicated to the vendor that the outgo would need to be appropriately adjusted for the monthly quota to be met. In the following months, the vendor increased outgo by approximately 4,000 to meet quota.

November Wave: The vendor MarketCube achieved its monthly quota of 450 completes a week before the close of the November wave due to an error in survey strategy and an improper application of quotas. Vendors are intended to achieve final completes during the last week of the wave for appropriate distribution of completes. Ipsos invalidated 20 completes from November 20 and November 21 to ensure that the vendor achieved completes during the final week of the wave. Ipsos requested that MarketCube set quotas accordingly and monitor their sample outgo in assurance that this does not become a recurring issue.

December Wave : During arrangements for the launch of the December wave, iSay failed to prepare and release their sample. This fault impacted three days of sampling for the vendor, and because the sample was not sent out, no completes were achieved through iSay for the first few days of the December wave. Once the error was discovered, sample outgo was marginally increased the following week to account for the three days of missing sample. Ipsos monitored the remainder of the wave to guarantee that the monthly quota of 1,500 completes was achieved without error. As this problem was due to human error, the issue was confirmed and escalated to upper management for further investigation. To ensure issues of this nature did not occur again, Ipsos implemented additional quality assurance checks on both ends of launching to verify that all vendors properly prepared their sample and released the sample on the first of the month.

Data Cleaning

Mismeasurement in survey responses can bias estimated stiffing and tipping rates. To mitigate bias, FMG applied several data cleaning procedures to the survey data prior to conducting analysis .

Repeat Respondents

Although individuals were prevented from responding to the survey multiple times in a given month, there was no procedure in place to prevent individuals from responding and repeating the survey in different months. Because an individual's tipping behavior over time may be more similar than the tipping behavior of two different individuals over time, the responses of repeat respondents across survey completes might not be independent, which can complicate statistical inference.

In addition, prior research on consumer panel surveys has shown that exposure to the survey instrument or the completion of a survey may influence respondent spending and saving behavior. [2] Consequently, individuals who already responded to a survey may no longer be representative of the wider population of interest with respect to tipping behavior.

Table 5: Number of Respondents by Number of Completed Surveys

Number of Completed Surveys	Number of Respondents
1	38,156
2	2,143
3	403
4	143
5	39
6	11
7	4

[2] Crossley, T. F., Bresser, J., Delaney, L., & Winter, J. (2017). Can survey participation alter household saving behaviour?. *The Economic Journal* .

Total	40,899

To mitigate these issues, for individuals who completed the survey multiple times, only data from the first completed survey were retained for analysis. Individual surveys were assigned to respondents using a survey-to-respondent crosswalk that was provided by the data vendor.[3] Of the 40,899 respondents, 2,743 (6.7%) had more than one completed survey (Table 5). A total of 3,613 out of 44,512[4] total completes (8.1% of the total) were dropped as a result of this procedure.

Repeated Transactions

There was also evidence that within a given completed survey, respondents were reporting the same transaction multiple times. Specifically, duplicate transactions for a given day were identified based on the tipping industry, bill size, and tip amount. These duplicate transactions might have reflected confusion on the part of the respondent with respect to the survey instructions, whereby respondents might have been unsure about the requested recall period, and thus, might have recorded transactions that took place over multiple days. Alternatively, respondents might have been confused as to whether the information about a given transaction was actually recorded, and thus, decided to enter it again.

To mitigate potential bias that resulted from these duplicate entries, for a set of transactions reported by a given respondent for a given day with the same subindustry, bill size, and tip amount, only one transaction was retained. A total of 2,320 out of 52,002 total transactions (4.5% of the total) were dropped as a result of this deduplication.

Detection of Extreme Values

We observed several instances of extremely high bill amounts, tip amounts, and tip rates in the survey data. Assuming some of the unusual and unexpected data points represent measurement error or invalid transactions, an outlier identification strategy—similar to that employed by FMG in the IRS Tipping Task Order 3 report, *Comparison of Estimates of Tipping Behavior Produced Using Probability and Non-Probability Samples: Methodology and Results*—was implemented for the current study.

Specifically, we assumed that total daily expenditure on bills and tip amounts were log normally distributed and tip rate was normally distributed for each transaction type (e.g., full-service restaurants, hairdressers).[5] Total bill and tip expenditures were used to identify outliers rather than

[3] Individual surveys were identified by Month, UniqueID, and Start Date. Individual respondents were identified by individual IDs that were specific to online panels. Individuals who participated in multiple online panels could not be identified. Consequently, some individuals may still be responsible for more than one survey in the estimation sample.

[4] This number excludes 341 completes that were classified as "abandoned," "error," or "quota full client."

[5] We recognize the normality assumption applied may not hold due to non-independence of transactions within commuting zones as well as individual respondents. However, the small number of transactions per commuting zone and individual makes identifying outliers by commuting zone and individual unfeasible. It should also be noted that the standard errors do not account for the identification of outliers. Under a different sample, the threshold for identifying outliers would be different, resulting in potentially significantly different estimates. This, along with uncertainty surrounding missing data for

characteristics of individual transactions , because the expenditures combine information on transaction frequency and transaction characteristics, both of which are necessary for calculating total tipped expenditure or transaction weighted stiffing and tipping rates. We then calculated the following ratio for each outcome by transaction type as follows:

$$\frac{|y - Q_{75}|}{Q_{75} - Q_{25}} \text{ for } y > Q_{75}$$

$$\frac{|y - Q_{25}|}{Q_{75} - Q_{25}} \text{ for } y < Q_{25}$$

In this case, *y* was logged total daily bill for a given service, logged total daily tips given to a type of service, or the ratio of total daily tips over total daily bills. Respondents were identified as outliers if, for a given transaction type, either of the above ratios exceeded 2.5 for bill amount, tip amount, or tip-to-bill ratio.[6] Respondents with at least one outlier transaction type were excluded from the analysis. A total of 6,467 out of 34,432 remaining respondents (15. 8%) were identified as outliers using this procedure. The sample remaining after these exclusions consisted of 34,432 respondents.

Weighting and Analysis

Given survey nonresponse as well as the potential for systematic differences between respondents determined to have provided outlier information and other respondents , the final set of respondents might not be representative of the population with respect to characteristics relevant to tipping behavior. This lack of representation may, in turn, result in biased estimates of average tip rates and stiffing rates.

Poststratification Weights

To mitigate such bias , poststratification weights were developed to allow estimates of key outcomes on a weighted sample that is representative of the target population of American adults with respect to observable characteristics that may be relevant to tipping behavior. Poststratification weights can be used in such circumstances to calculate weighted averages, in which greater weight is given to respondents whose characteristics are underrepresented in the sample relative to the population of interest, which in turn , reduces estimate bias.

To calculate post stratification weights, a simple raking algorithm created by FMG internally was used and run in Stata software. Initially, each respondent was given equal weights (i.e., values of 1). The algorithm compared the distribution of respondents across categories of one characteristic, such as age, to the distribution of the target population. The respondents' initial weights were adjusted by multiplying them by the ratio of the fraction of the population in the respondent's category to the fraction of the sample in the respondent's category. The process was then repeated for another

---

certain transaction characteristics, means that the resulting point and standard error estimates could be sensitive to minor changes in methodology, particularly for industries with smaller numbers of transactions.

[6] For a given variable, a 2.5 interquartile range (IQR) threshold would only identify approximately 0.005% of respondents as outliers under a normal distribution, and is thus a relatively conservative threshold.

variable, using the adjusted weights from the previous round rather than the initial weights to calculate the fraction of the sample in a given category. The process was replicated for all relevant variables, and then another cycle through each variable was initiated using the adjusted weights from the previous cycle. The variable order for a given replicate was as follows: Rural-Urban Continuum Codes, region, educational attainment, gender, age, race/ethnicity, day of the week, and month. There were a total of 10 replications. The raking algorithm ensured that the final weighted distribution of each variable used to rake the sample was very close to its distribution in the population. Descriptive statistics for sample (unweighted) and population (weighted) characteristics using these weights as well as target proportions are presented in Table F2 (in Appendix F). The population characteristics are identical to the targets used in the raking algorithm. Finally, the weights were scaled such that they summed to the product of the population of individuals 18 or older in 2015[7] by 365 days to facilitate the calculation of estimates of annual totals of tipped expenditure. Specifically, the total number of 2015 person days (number of individuals aged 18 or older in the United States as of the 2015 ACS*365) were allocated across respondents proportional to their non-scaled weight. Thus, it is a constant transformation (weight proportion*242,831,196*365). This was done to facilitate the calculation of total tips and expenditures on an annual basis and provides identical estimates with respect to scale-invariant statistics (e.g. means). The use of these weights is recommended when generating population estimates using the survey data. Although the preferred estimation sample comprises the set of responses with poststratification weights created by FMG, the final data set includes all completed surveys along with indicator variables to identify respondents who were not assigned a poststratification weight due to extreme expenditure and tipping behavior in a given industry.

Estimates for total tipped expenditure or averages of transaction characteristics, such as the tip rate or stiff rate can be estimated by calculating the weighted means or totals of these sample characteristics. FMG used Stata's *reg* (OLS) command (constant only model) to calculate means of the number of daily transactions, bill size, tipped expenditure, fractions of transactions with a cash bill, fractions of tipped transaction where the tip was paid in cash, and the fraction of transactions which were stiffed. The *total* command to calculate total tipped expenditure. For all variables, the post-stratification weights were treated as probability weights (*pweight* option). Respondents in the same commuting zone may have visited the same establishments. Consequently, a random pair of transactions from the same commuting zone may be more similar with respect to tip rate and stiffing probability than a random pair that took place in two different commuting zones. Standard errors should account for dependence within commuting zones and were calculated using the cluster option in the *reg* and *total* commands.

Estimates for population average tipping outcomes using the first six months of data collection are presented in the report titled *Interim Report on the Survey of Consumer Tipping Behavior* in Appendix E.

---

[7] The year 2015 was used because it was the last year for which the five-year American Community Survey (ACS) estimates of the population were available.

Multilevel Regression and Poststratification (MRP)

The IRS intends to use the consumer tipping data from this survey in a number of ways. One of those ways will be to develop subnational, industry-specific tipping rates. This section provides a discussion of how FMG developed those rates from the survey data.

One means of obtaining both nationally and subnationally representative estimates of tipping and stiffing rates is MRP (Gelman & Little, 1997;[8] see Buttice and Highton, 2013,[9] and Toshkov, 2015,[10] for recent reviews and critiques). Model-based poststratification strategies have been employed to generate estimates that conform to administrative data using non-representative samples.[11] MRP has attained popularity among social scientists who wish to obtain geographically disaggregated estimates of a quantity of interest. Awareness of variation in tipping rates faced by establishments in different parts of the country will be of potential use to the IRS in so far as it provides a general understanding of patterns of tipping behavior and it might help detect differences in compliance.

Analyzing consumer tipping data for a particular industry using MRP involves first estimating models of the number of transactions undertaken by consumers as well as their tipping behavior that take the form:

$$1)\ \hat{\square}_{ik}\square = \square^{\beta \square X_{ik} + \alpha \square G_k + C_{\square\square}}$$

$$2)\ \hat{\square}_{tik}\square = \frac{\square^{\beta \square X_{ik} + \alpha \square G_k + C_S}}{1 + \square^{\beta \square X_{ik} + \alpha \square G_k + C_S}}$$

$$3)\ \hat{\square}_{tik}\square \,|\, (\square_{tik}\square = 0) = \square^{\beta \square X_{ik} + \alpha \square G_k + C_{\square\square}}$$

$$4)\ \hat{\square}_{tik}\square \,|\, (\square_{tik}\square = 1) = \square^{\beta \square X_{ik} + \alpha \square G_k + C_{\square\square}}$$

$$5)\ \hat{T}_{tik} = \square^{\beta \square X_{ik} + \alpha \square G_k + C_T}$$

...in which $\hat{\square}_{ijk}\square$ is the expected total number of transactions engaged in by respondent *i* in location *k*; $\hat{\square}_{tik}\square$ is the expected probability that respondent's transaction *t* was tipped; $\hat{\square}_{tik}\square$ is the expected bill size for respondent's transaction *t*, which is allowed to vary based on whether or not it was tipped;[12] and $\hat{T}_{tik}$ is an expected tip rate for transaction *t* calculated by dividing a reported dollar amount in tips by transaction bill size; *X* is a set of observable respondent-level demographic variables that includes age, gender, and educational attainment, and that are likely to be correlated with both tipping behavior and the number of transactions; and *G* is a set of location-specific factors that include: the racial composition of the respondent's county (i.e., percentage Black, Hispanic, and

---

[8] Gelman, A., & Little, T. C. (1997). Post-stratification into many categories using hierarchical logistic regression. *Survey Methodology, 23* (2): 127–135.

[9] Buttice, M. K., & Highton, B. (2013). How does multilevel regression and post-stratification perform with conventional national surveys?. *Political Analysis*, 21(4), 449–467.

[10] Toshkov, D. (2015). Exploring the performance of multilevel modeling and post-stratification with eurobarometer data. *Political Analysis*, mpv009.

[11] Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting*, *31*(3), 980–991.
Goel, S., Obeng, A., & Rothschild, D. Non-representative surveys: Fast, cheap, and mostly accurate. Working Paper.

[12] The exception for this is full-service restaurants, for which only one average bill size is calculated, due to the small fraction of transactions that were not tipped.

Other); the percentage of the adult population that is foreign born; the fraction of households in the respondent's county in a given income bracket/ median household income of the county; size of respondent's metropolitan area/ whether the respondent is residing within a metropolitan area; and census region. These variables are intended to capture variability in the number of transactions and tipping behavior by sector that is not explained by differences in *X* between locations. See Table F3 for sample and population proportions for all predictors. The fraction of the county's population which are foreign born as well as the race/ ethnicity shares are treated as continuous variables while all other variables are treated as categorical. Note that while the location *k* is the most narrowly defined geographic area for which data is available, predictions can be generated for aggregated levels of geography *g*. Finally, *C* is a constant.

Table 6: Outcomes for Multilevel Regressions

Outcome	Definition	Variable Name*	Model/ Stata Command
Number of Daily Transactions ($\square_{ik}$)	Number of transactions of a given type paid for by respondents in the 24 -hour period before the survey.	N/ A	Poisson
Bill Size ($\square_{tik}$)	Amount of non -tip expenditure on a bill (e.g., sum of relevant menu prices).	T1_Q1_E	Poisson
Tipped Transaction ($\square_{tik}$)	One if transaction was tipped, "0" otherwise.	T1_Q1_F , T1_Q1_G	Logit
Tip Rate (T_{tik})	Ratio of tipped expenditure on transaction over non -tipped expenditure for a given transaction.	T1_Q1_F_Open , T1_Q1_I , T1_Q1_E	Poisson

*Variable name in transaction file .

Parameters β, α, and *C*, and predictions of $\hat{\square}$, $\hat{\square}$, and $\hat{\square}$ are estimated via Poisson regression (*poisson* command in Stata), whereas parameters for □ are estimated using a logistic regression (*logit* command in Stata). Table G1 contains descriptions of the outcome variables along with variable names. The resulting models are used to generate predictions (*predict* command in Stata) for each outcome for each strata defined by all *N* combinations of values of *X* and *G* covariates. Poststratification is then used to generate the predicted annual number of tipped transactions, transaction average stiffing rates, tipping rates, and ratios of total tipped expenditure to total bill size for a given location:

$$6)\ \hat{\#} = \sum_{s}^{N} \hat{E}_s \hat{\square}_s P_s \cdot 365$$

$$7)\ \overline{(1 - \hat{\square})} = \sum_{s}^{N} \frac{\hat{E}_s P_s}{\sum_s^N \hat{E}_s P_s} (1 - \hat{\square}_{tik})$$

$$8)\ \bar{\square} = \sum_{s}^{N} \frac{\hat{E}_s \hat{\square}_s P_s}{\sum_s^N \hat{E}_s \hat{\square}_s P_s} \hat{T}_s$$

9) [illegible] = [illegible], where:

$$\overline{\square} = \sum_{s}^{N} \hat{E}_s \hat{\square}_s P_s \hat{\square}_{tik}(\square_{tik} = 1) \hat{T}_s$$

$$\overline{\square} \ldots$$

$$= \sum_{s}^{N} (\hat{E}_s \hat{\square}_s P_s \hat{\square}_{tik}(\square_{tik} = 1)) + (\hat{E}_s (\hat{1} - \square_s) P_s \hat{\square}_{tik}(\square_{tik} = 0))$$

P is the population of a given demographic/geographic stratum *s* in a given location *g*, taken from the 2015 five-year American Community Survey (ACS). Commuting zone-level geographic factors are used to model individuals' number of transactions and tipping behavior. Predictions are generated for the United States as a whole as well as for commuting zones. The preferred subnational geographic unit is the commuting zone. Commuting zones are more likely to encompass the customer base of a given establishment. Commuting zones have been used in recent, prominent studies to define the geographic extent of environmental determinants of social outcomes.[13] Commuting zones may act as a proxy for the typical geographic extent of respondents' daily travels, and thus the establishments they are likely to visit.

This MRP procedure was undertaken separately for each industry, excluding home maintenance, hotels, and casinos, where either there is a significant likelihood that the transaction took place outside the respondent's commuting zone of residence or the number of transactions is extremely low. To quantify the uncertainty in the estimates that results from sampling variability, a cluster bootstrap procedure was used. Specifically, 1,000 samples of commuting zones were drawn with replacement (i.e. the sample commuting zone can enter multiple samples). For each replicate sample, the data from all respondents residing in the commuting zones represented in that replicate sample were used to generate the MRP estimates for transaction average stiffing rate, tipping rate, and tipping percentage. This resulted in 1,000 replicate estimates of the transaction average stiffing rate, tipping rate, and tipping percentage for each county or commuting zones. The standard deviation of the replications is the standard error of the estimate. A separate table provided in the interim report included estimates and standard errors for each commuting zone based on the first six months of data. The step-by-step MRP procedures are outlined in the Table 7.

[13] Chetty, R., Hendren, N., Kline, P., & Saez, E. (2014). Where is the land of opportunity? The geography of intergenerational mobility in the United States. *The Quarterly Journal of Economics*, *129*(4), 1,553–1,623.

Table 7: MRP Estimation Steps

Step	Description
	Point Estimates
1	Open Transaction File
2	Estimate Equations 1 -5 using all transactions with weight
3	Open Frame file with one row per age -gender-education-county stratum with P of each stratum
4	Using Parameters from Step 2 and P of each stratum, generate predictions for mean □$_{ik}$, □□$_{tik}$□∣ S=1 , □$_{tik}$□∣ S=0, □$_{tik}$, □ and T_{tik} for each age-gender-education-county stratum
5	Using predictions from Step 4, generate aggregate outcomes in Equations 6 - 9 for each Commuting Zone
	Standard Errors
6	Open Transaction File
7	Cluster sample* by commuting zones with replacement respondents/ transactions
8	Using all respondents/ transactions with weights drawn in Step 7, Estimate Equations 1 -5
9	Open Frame with one row per age -gender-education-county stratum with P of each stratum
10	Using Parameters from Step 8 and P of each stratum, generate predictions for mean □$_{ik}$, □□$_{tik}$□∣ S=1 , □$_{tik}$□∣ S=0, □$_{tik}$, □ and T_{tik} for each age-gender-education-county stratum
11	Using predictions from Step 10, generate aggregate outcomes in Equations 6-9 for each Commuting Zone
12	Repeat Steps 6-11 999 more times, resulting in 1,000 separate estimates (in addition to the point estimates) for Equations 6 -9
13	Take standard deviation of 1000 estimates for Equations 6 -9 to obtain bootstrapped standard errors fo r commuting zone outcomes described in Equation 6-9

*bsample, cluster() command in Stata

Appendix A: Estimating Consumer Tipping Behavior: Review and Recommendations (February 2014)

Estimating Consumer
Tipping Behavior (Feb

Appendix B: IRS Tipping Report on Cognitive and Usability Testing

IRS Tipping Report
on Cognitive and Usal

Appendix C: Final 2017 Consumer Tipping Survey

2017 IRS Tipping Questionnaire.docx

Appendix D: Comparison of Estimates of Tipping Behavior Produced Using Probability and Non-Probability Samples: Methodology and Results

Comparison of
Probability and Non-P

Appendix E: Interim Report on the Survey of Consumer Tipping Behavior (January 2018)

Interim Report on
the Survey of Consum

Appendix F: Descriptive Statistics

Table F1: Age by Gender by Region by Income Balancing

Sampling Cell	Balancing %
Male 18–34 Northeast Under $20K	0.333
Male 18–34 Northeast $20K –$49.9K	0.675
Male 18–34 Northeast $50K –$99.9K	1.095
Male 18–34 Northeast $100K+	0.605
Male 18–34 Midwest Under $20K	0.442
Male 18–34 Midwest $20K –$49.9K	0.898
Male 18–34 Midwest $50K –$99.9K	1.219
Male 18–34 Midwest $100K+	0.562
Male 18–34 South Under $20K	0.79
Male 18–34 South $20K –$49.9K	1.604
Male 18–34 South $50K–$99.9K	1.985
Male 18–34 South $100K+	1.004
Male 18–34 West Under $20K	0.491
Male 18–34 West $20K–$49.9K	1.043
Male 18–34 West $50K–$99.9K	1.33
Male 18–34 West $100K+	0.655
Male 35–54 Northeast Under $20K	0.286
Male 35–54 Northeast $20K –$49.9K	0.825
Male 35–54 Northeast $50K –$99.9K	1.351
Male 35–54 Northeast $100K+	1.07
Male 35–54 Midwest Under $20K	0.34
Male 35–54 Midwest $20K –$49.9K	0.955
Male 35–54 Midwest $50K –$99.9K	1.486
Male 35–54 Midwest $100K+	0.863
Male 35–54 South Under $20K	0.646
Male 35–54 South $20K –$49.9K	1.641
Male 35–54 South $50K –$99.9K	2.453
Male 35–54 South $100K+	1.463
Male 35–54 West Under $20K	0.396
Male 35–54 West $20K–$49.9K	0.992
Male 35–54 West $50K–$99.9K	1.531
Male 35–54 West $100K+	1.011
Male 55+ Northeast Under $20K	0.351
Male 55+ Northeast $20K –$49.9K	1.091
Male 55+ Northeast $50K –$99.9K	0.991
Male 55+ Northeast $100K+	0.572
Male 55+ Midwest Under $20K	0.391
Male 55+ Midwest $20K –$49.9K	1.326

Sampling Cell	Balancing %
Male 55+ Midwest $50K –$99.9K	1.264
Male 55+ Midwest $100K+	0.608
Male 55+ South Under $20K	0.75
Male 55+ South $20K –$49.9K	2.166
Male 55+ South $50K –$99.9K	2.04
Male 55+ South $100K+	1.066
Male 55+ West Under $20K	0.448
Male 55+ West $20K –$49.9K	1.183
Male 55+ West $50K –$99.9K	1.174
Male 55+ West $100K+	0.626
Female 18–34 Northeast Under $20K	0.311
Female 18–34 Northeast $20K – $49.9K	0.656
Female 18–34 Northeast $50K – $99.9K	1.001
Female 18–34 Northeast $100K+	0.516
Female 18–34 Midwest Under $20K	0.415
Female 18–34 Midwest $20K –$49.9K	0.846
Female 18–34 Midwest $50K –$99.9K	1.335
Female 18–34 Midwest $100K+	0.565
Female 18–34 South Under $20K	0.745
Female 18–34 South $20K –$49.9K	1.5
Female 18–34 South $50K –$99.9K	2.352
Female 18–34 South $100K+	1.095
Female 18–34 West Under $20K	0.474
Female 18–34 West $20K –$49.9K	1.021
Female 18–34 West $50K –$99.9K	1.413
Female 18–34 West $100K+	0.662
Female 35–54 Northeast Under $20K	0.24
Female 35–54 Northeast $20K – $49.9K	0.784
Female 35–54 Northeast $50K – $99.9K	1.209
Female 35–54 Northeast $100K+	0.696
Female 35–54 Midwest Under $20K	0.295
Female 35–54 Midwest $20K –$49.9K	0.934
Female 35–54 Midwest $50K –$99.9K	1.714
Female 35–54 Midwest $100K+	0.915
Female 35–54 South Under $20K	0.555
Female 35–54 South $20K –$49.9K	1.625
Female 35–54 South $50K –$99.9K	3.054
Female 35–54 South $100K+	1.851
Female 35–54 West Under $20K	0.357

Sampling Cell	Balancing %
Female 35–54 West $20K–$49.9K	0.982
Female 35–54 West $50K–$99.9K	1.745
Female 35–54 West $100K+	1.162
Female 55+ Northeast Under $20K	0.389
Female 55+ Northeast $20K –$49.9K	1.385
Female 55+ Northeast $50K –$99.9K	1.203
Female 55+ Northeast $100K+	0.579
Female 55+ Midwest Under $20K	0.484
Female 55+ Midwest $20K –$49.9K	1.642
Female 55+ Midwest $50K –$99.9K	1.486
Female 55+ Midwest $100K+	0.632
Female 55+ South Under $20K	0.844
Female 55+ South $20K –$49.9K	2.683
Female 55+ South $50K –$99.9K	2.413
Female 55+ South $100K+	1.125
Female 55+ West Under $20K	0.462
Female 55+ West $20K –$49.9K	1.518
Female 55+ West $50K –$99.9K	1.373
Female 55+ West $100K+	0.672

Table F2: Poststratification Variables, Weighted and Unweighted Proportions (Sources for Target Proportions, N = 34,170)

Category	Population Target	Unweighted Proportion	Weighted Proportion
	Rural–Urban Continuum Codes (ACS 5-Year)		
Counties in metro areas of 1 million population or more	54.90%	53.91%	54.90%
Counties in metro areas of 250,000 to 1 million population	21.17%	22.19%	21.17%
Counties in metro areas of fewer than 250,000 population	9.18%	9.74%	9.18%
Non-metro counties	14.75%	14.16%	14.75%
	Census Region (ACS 5-Year)		
Northeast	18.09%	21.34%	18.09%
Midwest	21.31%	24.09%	21.31%
South	37.30%	34.72%	37.30%
West	23.29%	19.85%	23.29%
	Educational Attainment (ACS 5-Year)		
High School or Less	41.50%	22.89%	41.50%
Some College	31.30%	37.13%	31.30%
College	17.31%	25.80%	17.31%
Graduate Degree	9.89%	14.18%	9.89%
	Gender (ACS 5-Year)		
Female	51.37%	60.25%	51.37%
Male	48.63%	39.75%	48.63%
	Age (ACS 5-Year)		
18–24	12.92%	7.56%	12.92%

Category	Population Target	Unweighted Proportion	Weighted Proportion
25–34	17.66%	18.46%	17.66%
35–44	16.74%	14.77%	16.74%
45–64	34.31%	39.72%	34.31%
65+	18.37%	19.50%	18.37%
	Race/Ethnicity (Ipsos Poststratification Weights)		
Non-Hispanic White	69.43%	80.64%	69.43%
Non-Hispanic Black	12.57%	7.37%	12.57%
Hispanic	9.87%	6.11%	9.87%
Other Non-Hispanic	8.13%	5.88%	8.13%
	Day of the Week of Survey Start Date		
Sunday	14.29%	12.51%	14.29%
Monday	14.29%	14.78%	14.29%
Tuesday	14.29%	14.32%	14.29%
Wednesday	14.29%	13.97%	14.29%
Thursday	14.29%	14.54%	14.29%
Friday	14.29%	16.63%	14.29%
Saturday	14.29%	13.24%	14.29%
	Month of Survey Start Date		
January	8.49%	11.61%	8.49%
February	7.67%	8.72%	7.67%
March	8.49%	8.73%	8.49%
April	8.22%	8.12%	8.22%
May	8.49%	8.21%	8.49%
June	8.22%	7.70%	8.22%
July	8.49%	7.66%	8.49%

Category	Population Target	Unweighted Proportion	Weighted Proportion
August	8.49%	7.63%	8.49%
September	8.22%	8.22%	8.22%
October	8.49%	7.78%	8.49%
November	8.22%	7.75%	8.22%
December	8.49%	7.86%	8.49%

Table F3: MRP Poststratification Variables, Population Proportions (Sources for Variables, *N* = 34,170)

Category	Sample Proportion	Population Proportion
Individual-Level Variables (X_{ik})		
Gender (DEM_3; Reference: Female)		
Male	39.75%	48.63%
Age (DEM_1; Reference: 18-24)		
25–34	18.46%	17.66%
35–44	14.77%	16.74%
45–64	39.72%	34.31%
65+	19.50%	18.37%
Educational Attainment (DEM_6_FINAL; Reference: High School or Less)		
Some College	37.13%	31.30%
College	25.80%	17.31%
Graduate Degree	14.18%	9.89%
County-Level Variables (G_k)		
% of County Foreign Born (Appended based on *FIPS*)	14.48%	16.13%
Race/Ethnic Composition (Appended based on *FIPS; Reference: % of County Population, White)*		
% of County Population, Black	11.73%	12.25%
% of County Population, Hispanic	14.81%	16.89%
% of County Population, Other	7.54%	8.29%
% of County Households by Income Bracket (DEM_8_FINAL; Reference: % of Households <$10,000)		
% of Households, $10,000 – $14,999	5.23%	5.25%
% of Households, $15,000 – $24,999	10.60%	10.54%
% of Households, $25,000 – $34,999	10.15%	10.05%
% of Households, $35,000 – $49,999	13.48%	13.36%
% of Households, $50,000 – $74,999	17.92%	17.78%
% of Households, $75,000 – $99,999	12.17%	12.12%
% of Households, $100,000 –	13.12%	13.20%

Category	Sample Proportion	Population Proportion
$149,999		
% of Households, $150,000+	10.19%	10.51%
Rural–Urban Continuum Codes (Appended based on FIPS; Reference: Counties in metro areas of 1 million population or more)		
Counties in metro areas of 250,000 to 1 million population	22.19%	21.17%
Counties in metro areas of fewer than 250,000 population	9.74%	9.18%
Non-metro counties	14.16%	14.75%
Census Region (Appended based on *FIPS; Reference: Northeast)*		
Midwest	24.09%	21.31%
South	34.72%	37.30%
West	19.85%	23.29%

Appendix G: Glossary of Terms

Table G1: Glossary of Terms

Outcome	Definition	Variable Name*
Number of Daily Transactions	Number of transactions of a given type paid for by respondents in the 24 -hour period before the survey.	N/ A
Bill Size	Amount of non -tip expenditure on a bill (e.g., sum of relevant menu prices).	T1_Q1_E
Cash Bill	Yes, if non-tipped expenditure was paid in cash, "0" otherwise.	T1_Q1_D_(1-8)
Tipped Expenditure	Expenditure for a given transaction that takes the form of a tip.	T1_Q1_F_Open , T1_Q1_I
Total Tipped Expenditure	Total tipped expenditure across all transactions of a given type.	T1_Q1_F_Open , T1_Q1_I
Stiff Rate	Percentage of transactions in which there was no tipped expenditure.	T1_Q1_F , T1_Q1_G
Tip Rate	Ratio of tipped expenditure on transaction over non-tipped expenditure for a given transaction .	T1_Q1_F_Open , T1_Q1_I , T1_Q1_E
Tipped Percentage	The ratio of the total tipped expenditure across all transactions of a given type to the total bill size across all transactions of that type.	T1_Q1_F_Open , T1_Q1_I , T1_Q1_E ,
Cash Tip	Yes, if tipped expenditure was paid in cash, "0" otherwise.	T1_Q1_H_(1-8)

*Variable name in transaction file .