
# Comparison of Estimates of Tipping Behavior Produced using Probability and Non-Probability Samples

Prepared for Internal Revenue Service

Prepared by Fors Marsh Group LLC

January 2015

Version 1.1

## Table of Contents

## Introduction

The potential target population for the IRS tipping study includes all U.S. resident persons who use services that are commonly tipped. A precise estimate of 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 not to 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. 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 the current study.

With respect to lowering the cost of data collection, an increasingly common alternative is the use of non-probability 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 non-response associated with pushing individuals sampled from the telephone or mail frame to the Internet survey instrument. 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, 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. 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.

[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).

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 will follow Office of Management and Budget (OMB) guidelines[2] by using a pilot survey to resolve this conflict. Specifically, we will conduct a pilot study to determine if the results generated by two different Internet-based data streams—a probability-based sample derived from the GfK KnowledgePanel and a non-probability based sample taken from Ispos' i-Say online opt-in panel—produce equivalent estimates. This will allow 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 facilitate reduced respondent burden, lower item non-response rates, and greater response accuracy than mail- or phone-based surveys.

Non-probability Based Sample: 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.

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. 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 error. This feature, along with the use of a web-based interface, would allow for the calculation of unbiased estimates of tipping behavior from a probability-based sample. Consequently, if estimates derived from the Ipsos and GfK samples support identical conclusions about the tipping behavior across industries and geographic areas, this would lend support to the more cost-efficient non-probability based method. In this event, the use of the i-Say panel would generate more cases at lower cost per case than would be the case with a probability-based sample, without a substantial cost 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.

[2] 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."

## Methodology

This section describes two methodologies that can be used to decide between the use of probability and non-probability samples for the final 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, under the assumption that the non-probability sample is at least as biased with respect to population tip rates as the probability sample and less costly per complete. 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, under the assumption 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 with respect to probability-based panels because of its combination of a representative frame. Under the assumption that an estimate derived from a probability sample is at least as unbiased 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, assuming equal variability in tipping rate between the populations represented in the two samples, reduced to a bias versus variance trade-off. Given that it is known that the cost-per complete will be lower with the non-probability sample, then if the samples do not differ with respect to tipping behavior, 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 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 separately for each industry:

$$1)\hat{T}_{tj} = \delta Ipsos_{tj} + Constant$$

In Equation 1, $\hat{T}_{tj}$ is a tip rate greater than 0 of positive tip, positive bill size transaction t for an individual residing in location j; $Ipsos$ 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 either differences in observed or unobserved demographic or geographic characteristics of respondents in the two samples. Specifically, a $\delta$ 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 parameters ($k$=2) of this model, it allows for precise estimates of this unconditional difference even with small samples. Given that some industries are likely to have a small number of tipping transactions represented in the data, the parsimony of Equation 1 becomes a big advantage. Note that the assumption that the variance of the two samples is constant can be tested using standard tests for heteroscedasticity (e.g. Breusch-Pagan, Brown-Forsythe). The test for bias in the in the non-probability sample can be made robust to violation of equal variances through the use of robust standard errors.

However, the Pilot Survey is expected to field to approximately 20,000 respondents, with the GfK KnowledgePanel and Ipsos i-Say panel each contributing approximately 10,000. Consequently, the data may support the estimation of more complex models, especially for those industries where potentially tipped transactions are more frequent. In particular, full-service restaurants are expected to be well represented in the sample of tipped transactions based on the analysis of tip frequency

presented in the final report for Task Order #1.[3] Given, as discussed in the next section, the "Differences in Samples" test can only be applied to full-service restaurants, the recommendation with respect to the use of a probability or non-probability sample may ultimately be determined by the results for this sector. Consequently, there may be little cost in terms of the scope of the analysis in using the more complex model. In addition, given that we can weight the sample that results from the final fielding to match the demographic and geographic characteristics of our population of interest, the IRS may not be interested in differences in tipping behavior between the two samples explained by differences in observable demographic characteristics. Consequently, we may instead 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 for each industry:

$$2)\hat{T}_{tj} = \delta Ipsos_{tj} + \beta X_{tj} + \alpha G_j + Constant$$

In Equation 2, $X$ is a vector of demographic characteristics observable in both samples as well as in our frame (CPS, American Community Survey (ACS), etc.); and $G$ is a vector of geographic characteristics, where the geographic unit could be counties, commuting zones, or Direct Marketing Areas (DMAs). If parameter $\delta$ is significantly different from 0[4], then the estimated model is consistent with a conditional difference in tipping rates between the two samples.

### "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 no more biased than 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 are inclined to join opt-in Internet panels (e.g., i-Say panelists) do not conditionally or unconditionally differ from the general population with respect to tipping, but those inclined to respond to solicitations through the mail, and thus participate in GfK's KnowledgePanel, do. To examine whether the conclusions drawn from the first part of the analysis are robust to relaxing this assumption, probability and non-probability estimates of tipping rates are compared with estimates derived from POS data.

The primary assumption of this part of the analysis is that the POS data is relatively unbiased as an 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. However, this assumption may be violated if there is systematic misrecording 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 document *An Assessment of the Validity of Using Point-of-Sale Data to Estimate Restaurant Tipping Rates*[5] discussed 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. Both types of transactions may differ from non-cash, non-prepayment transactions with respect to their "true" tipping rates. Consequently, using the POS data as a benchmark will likely only be valid for non-cash, non-prepayment transactions. The POS validation report also found issues with respect to establishment "non-response." Specifically, there were too few quick-service tipping

[3] *Estimating Consumer Tipping Behavior: Review and Recommendations* (2014). Internal report prepared for the Internal Revenue Service by Fors Marsh Group under contract TIRNO-13-Z-00021-001.
[4] The analysis will use the standard .05 threshold for statistical significance.
[5] *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.

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. This meant that the POS data can only be used as a baseline for full-service restaurants. And although the report found little evidence of systematic differences in establishment representation across DMAs, there was no ability to test for differential establishment inclusion within DMAs. These issues may undermine the reliability of the POS-derived estimates. Consequently, this "Differences in Differences" analysis does not strictly dominate the "Differences in Sample" analysis.

To estimate the unconditional "Differences in Differences," we can pool data for tipped transactions at full-service restaurants from the probability, non-probability, and POS samples (or a random subsample of the latter to mitigate computational complexity) and estimate the following model separately for each industry:

$$3)\ \hat{T}_{tj} = \delta Ipsos_{tj} + \vartheta GfK_{tj} + \ Constant$$

Equation 3 differs from Equation 1 in that it includes $\mathrm{GfK}_{tj}$, an indicator variable that takes a value of 1 if a given transaction was extracted from the GfK sample and 0 otherwise. The excluded category is now transactions taken from the POS sample. Our null hypothesis can be stated as:

$$|\delta| = |\vartheta|$$

This null hypothesis can be tested using a Wald or Likelihood Ratio Test. If the null hypothesis is rejected, we may conclude that the survey sample with the smaller absolute difference more closely matches the mean tip rate implied by the POS data.

Similarly, the analogue of Equation 2 could be written as:

$$4)\hat{T}_{tj} = \delta Ipsos_{tj} + \vartheta GfK_{tj} + \ \alpha G_j + Constant$$

Note, however, that, unlike Equation 2, Equation 4 does not include controls for individual covariates. This is due to the fact that we do not observe individuals in the POS data. Consequently, differences in $\delta$ and $\vartheta$ may reflect differences in response across demographic groups within the same geographic units between surveys. This is problematic insofar as it is likely that the survey samples can be subjected to poststratification based on these demographic groups in the final analysis. Equation 4 may consequently lead to incorrect inferences about the true "Differences in Differences." As an alternative, we may estimate the following conditional models separately for each industry:

$$5a)\hat{T}_{tj} - \ \hat{\bar{T}}_{jPOS} = \delta Ipsos_{tj} + \beta X_{tj} + \ \alpha G_j + Constant$$

$$5b)|\hat{T}_{tj} - \ \hat{\bar{T}}_{jPOS}| = \delta Ipsos_{tj} + \beta X_{tj} + \ \alpha G_j + Constant$$

The left-hand side of the equation is now the deviation of a survey transaction tip rate from the estimated average tip rate implied by the POS average ($\hat{\bar{\mathrm{T}}}_{\mathrm{jPOS}}$) for the transaction's geographic unit. Controlling for the geographic average tipping rate for the POS transactions by subtracting it from the left-hand side removes all potential explanatory power for the geographic attributes of the POS transactions. Consequently, the POS transactions can be dropped from the analysis. Restricting the sample to survey transactions allows for the incorporation of respondent-level predictors. Note, however, that the interpretation of the regression coefficients now changes to the marginal effects of the predictors on the difference between a survey respondent from the expected tip of the

geographic unit rate as indicated by POS data. We can thus interpret Equations 5a and 5b as models of within geographic unit selection bias if we assume the POS data as the gold standard. Note that Equation 5a allows for a test of differences in the systematic deviation of respondents between samples in the same direction across geographic units, while 5b allows the direction of the deviations to vary across geographic units. Consequently, Equation 5a may be more useful for determining relative bias of the panels for the national mean tipping rate, and 5b may be more useful for testing for relative bias at the local level, which is relevant to the extent that the IRS may eventually want to develop small area estimates of tipping rates. The parameters of Equation 5b may also be interpreted as reflecting the differences in the degree of dispersion around the local area average tip rate between different samples and strata, and thus can be used to locate sources of variability (and thus potentially unreliability) in the different samples.

Note, however, that $\hat{\bar{T}}_{jPOS}$ is at the very least subject to sampling error. We might consequently consider weighting the regression by the inverse of the standard error of $\hat{\bar{T}}_{jPOS}$, in keeping with the methodology used to generate DMA-level mean tip rates in the earlier report assessing the validity of the POS data.[6] It would also be advisable to cluster standard errors at the level of the geographic unit to account for the automatic correlation in residuals that the inclusion of $\hat{\bar{T}}_{jPOS}$ on the left hand side induces across units in the same geographic unit.

The null hypothesis then becomes:

$$6)\ \left|E\left(\hat{T}_{tj} - \hat{\bar{T}}_{jPOS}\middle|Ipsos_{tj}=1\right)\right| = \left|E\left(\hat{T}_{tj} - \hat{\bar{T}}_{jPOS}\middle|Ipsos_{tj}=0\right)\right|$$

, which can be tested using a Wald or Likelihood Ratio Test. 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.

### 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 these results in such a way that 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 (disaggregating by industry; tip rate versus conditional versus unconditional tests; using weights or population data from Census/ACS for post-stratification) of these "Differences in Samples" and "Differences in Differences" tests are likely to be implemented for the purpose of robustness.

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 recommend using the probability sample for the full fielding; otherwise, the FMG Team will recommend the use of the non-probability sample. The first part of this rule is a result of the continued skepticism of non-probability samples among many survey statisticians[7]. In addition, there is a potential lack of external validity of tests utilizing the POS

[6] *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.
[7] 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

data with respect to bias in reported tipping transactions from establishments other than full-service restaurants where the bill or tip was paid non-electronically. The second part of this rule is based on the assumed lower cost of 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 – Potential Decision Matrix

| | | "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 beliefs, 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 – Potential Decision Matrix Adjusted for Alternative Set of Beliefs

| | | "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.

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 aside 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.

## Summary and Conclusions

This 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, these methodologies 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 "Differences in Differences" test does not make assumptions about the relative bias in the two samples, it can only be used to assess the reliability of reported tip rates for transactions undertaken at full-service restaurants using electronic payment methods. By contrast, the "Differences in Sample" test allows for an assessment in bias in reported stiffing rates in addition to tip rates and also allows comparisons in bias across different industries. Given these trade-offs between the two tests, assumptions will have to be made in the event that the two tests differ with respect to which sample their results imply should be used for the final survey.