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 Ispos'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.

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.

---

[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:

$$1)\ \hat{T}_{tijs} = \delta Ipsos_s + Constant$$

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:

$$2)\ \hat{T}_{tijs} = \delta Ipsos_s + \beta X_i + \alpha G_j + Constant$$

In Equation 2, X_i 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 G_j 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 β or α 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 β and α 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:

$$3a)\hat{T}_{tijs} - \hat{\bar{T}}_{jPOS} = \delta Ipsos_s + Constant$$

[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.

$$3b)\ |\hat{T}_{tijs} - \hat{\bar{T}}_{jPOS}| = \delta Ipsos_s + Constant$$

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

$$4a)\ \hat{T}_{tijs} - \hat{\bar{T}}_{jPOS} = \delta Ipsos_s + \beta X_i + \alpha G_j + Constant$$

$$4b)\ |\hat{T}_{tijs} - \hat{\bar{T}}_{jPOS}| = \delta Ipsos_s + \beta X_i + \alpha G_j + Constant$$

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 ($\hat{\bar{T}}_{jPOS}$) 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 estimate 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 $\hat{\bar{T}}_{jPOS}$ 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 $\hat{\bar{T}}_{jPOS}$ on the left hand side induces across units in the same commuting zone due to the use of the same/similar businesses and other local area characteristics.

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

$$5)\ \left|E\left(\hat{T}_{tijs} - \hat{\bar{T}}_{jPOS}\middle| Ipsos_s = 1\right)\right| = \left|E\left(\hat{T}_{tijs} - \hat{\bar{T}}_{jPOS}\middle| Ipsos_s = 0\right)\right|$$

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 evaluated using the well-known Wald 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 rate 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 approach.

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 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 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 – 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 (δ) 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
δ	-0.004	-0.006	-0.003	-0.005	0.003	0.006
	(-0.003)	(0.003)*	(-0.003)	(-0.003)	(-0.002)	(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 (δ) 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
δ	-0.002	-0.004	-0.001	-0.003	0.003	0.005
	(-0.003)	(-0.003)	(-0.004)	(-0.004)	(-0.002)	(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, an outlier identification strategy similar 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:

$$\frac{|y - y_{75thPercentile}|}{y_{75th\ Percentile} - y_{25th\ Percentile}} \text{ for } y > y_{75th\ Percentile}$$

$$\frac{|y - y_{25thPercentile}|}{y_{75th\ Percentile} - y_{25th\ Percentile}} \text{ for } y < y_{25th\ Percentile}$$

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 the full service restaurant transactions reported by these excluded individuals are reported separately for GfK and Ipsos respondents in Tables 7 and 8.

Descriptive Statistics

Table 7 – Descriptive Statistics for Outlying Full Service Restaurant Transactions - 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 (0.005)		-0.010 (0.007)		0.003 (0.003)
West North Central		-0.015 (0.007)*		-0.022 (0.008)**		0.012 (0.005)**
South Atlantic		-0.004 (0.005)		-0.025 (0.007)**		0.016 (0.003)**
East South Central		-0.018 (0.007)*		-0.029 (0.012)*		0.014 (0.007)
West South Central		-0.012 (0.006)*		-0.032 (0.008)**		0.025 (0.003)**
Mountain		-0.015 (0.005)**		-0.029 (0.008)**		0.016 (0.004)**
Pacific		-0.013 (0.007)		-0.007 (0.009)		0.005 (0.004)
Constant	0.184 (0.002)**	0.170 (0.018)**	-0.032 (0.003)**	-0.044 (0.020)*	0.052 (0.002)**	0.083 (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 (0.002)	0.185 (0.002)	-0.032 (0.003)	-0.030 (0.002)	0.052 (0.002)	0.051 (0.001)
Ipsos Predicted Value	0.180 (0.002)	0.179 (0.002)	-0.034 (0.003)	-0.035 (0.002)	0.056 (0.002)	0.057 (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 (0.003)	-0.004 (0.003)	-0.001 (0.004)	-0.003 (0.003)	0.003 -0.002	0.005 (0.002)*
Male		0.000 (0.003)		0.000 (0.003)		0.002 (0.003)
Age, 25-34		-0.018 (0.009)*		-0.019 (0.009)*		0.010 (0.005)
Age, 35-44		-0.024 (0.009)**		-0.023 (0.009)*		0.005 (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 (.017)**	.281 (.019)**	1.017 (0.093)**	1.259 (0.100)**
Some College	.376 (.019)**	.252 (.023)**	1.719 (0.106)**	1.137 (0.115)**
Associate Degree	.384 (.021)**	.256 (.024)**	1.758 (0.108)**	1.157 (0.119)**
Bachelor's Degree	.432 (.019)**	.299 (.022)**	1.973 (0.104)**	1.337 (0.110)**
Graduate Degree	.421 (.021)**	.289 (.025)**	1.926 (0.112)**	1.296 (0.121)**
Black	-.119 (.016)**	-.051 (.018)**	-0.528 (0.072)**	-0.217 (0.079)**
Hispanic	-.065 (.016)**	-.010 (.017)	-0.289 (0.071)**	-0.043 (0.074)
Other	-.018 (.023)	-.029 (.034)	-0.083 (0.100)	-0.125 (0.142)
Income, $10k-$14.9k	-.001 (.000)**	-.001 (.000)*	-0.003 (0.001)**	-0.003 (0.001)*
Income, $15k-$24.9k	.000 (.000)	.000 (.000)	-0.001 (0.001)	0.001 (0.001)
Income, $25k-$34.9k	-.001 (.000)**	-.001 (.000)**	-0.004 (0.001)**	-0.003 (0.001)**
Income, $35k-$49.9k	-.001 (.000)**	-.001 (.000)**	-0.004 (0.001)**	-0.004 (0.001)**
Income, $50k-$74.9k	-.001 (.000)**	-.001 (.000)**	-0.006 (0.001)**	-0.005 (0.001)**
Income, $75k-$99.9k	-.002 (.000)**	-.002 (.000)**	-0.009 (0.001)**	-0.010 (0.001)**
Income, $100k-$149.9k	-.003 (.000)**	-.002 (.000)**	-0.012 (0.001)**	-0.008 (0.001)**
Income, $150k+	-.003 (.000)**	-.002 (.000)**	-0.013 (0.001)**	-0.008 (0.001)**
Foreign Born, % of County Population	.160 (.056)**	.129 (.063)*	0.704 (0.252)**	0.541 (0.263)*
Metro Population, 250k - 1 Million	-.019 (.012)	-.018 (.014)	-0.084 (0.054)	-0.076 (0.060)
Metro Population, <250k	-.016 (.017)	-.022 (.020)	-0.070 (0.075)	-0.094 (0.083)
Non-Metro County	-.024 (.016)	-.022 (.018)	-0.106 (0.070)	-0.093 (0.075)
Middle Atlantic	.039 (.023)	.021 (.027)	0.172 (0.102)	0.091 (0.114)
Midwest	.028 (.024)	.047 (.029)	0.121 (0.106)	0.198 (0.123)
West North Central	-.028 (.029)	-.017 (.035)	-0.127 (0.127)	-0.079 (0.149)
South Atlantic	.022 (.022)	.048 (.027)	0.099 (0.097)	0.202 (0.113)
East South Central	.009 (.026)	.022 (.030)	0.044 (0.112)	0.092 (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