Submitted for publication to *Managing Service Quality*

Survey Mode Impact Upon Responses and Net Promoter Scores

Dr. Frederick C. Van Bennekom, Northeastern University
Samuel Klaidman, Middlesex Consulting

Abstract

Customer surveys are a very common method used by companies to gather feedback from customers. However, the validity of survey results can be compromised as a result of many biases that can be introduced into the data from the instrument design and the administration procedures. This research study examines the impact of survey mode, specifically telephone versus web-form survey modes, using actual data from a company that serves business clients. Responses for telephone surveys were found to be significantly higher than for web-form surveys, particularly at the top end of the response scale. This tendency has been seen in previous research, and we suggest this may result from a scale truncation effect.

Many companies today are using a new statistical measure, "net scoring," as a summary measure of a survey question's data. Specifically, net scoring has been applied to the recommendation question to arrive at a Net Promoter Score, which Fred Reichheld has shown to be the best single indicator of a company's future profitability. However, the net scoring procedure has serious threshold effects. Changes in composition of a mixed-mode survey program can result in fluctuating Net Promoter Scores that do not reflect changes in perceptions of the customer base, but rather are measurement errors introduced by the survey practices. More generally, our research shows the dangers of comparing survey scores across companies where surveying practices, including mode, may differ.

Introduction

Surveys are a widely used measurement tool of customer sentiment, though surveys as a research method have many validity issues. These issues fall mainly into the two categories of instrumentation bias and various administration biases. While the professional research community recognizes that validity issues exist with survey research, validity issues are neither well understood nor considered in the business community where surveying to capture customer feedback has expanded greatly over the past few decades, especially with the advent of web-form surveying tools.

So long as the data are being used only within an organization for problem identification and performance trending purposes, that is, examining changes in the perceptions from an organization's stakeholder base over time for a given survey, these validity issues are not a serious concern. The trend is reasonably legitimate unless the biases are particularly severe and change across administrations of the survey instrument. However, the validity issues are a real concern when surveying practices change within an organization or with cross-organization comparisons of survey data generated by surveys unique for each organization.

This paper presents research regarding one type of administration bias, the mode bias, to show

the impact that different modes of survey administration can have upon survey data, specifically telephone versus web-form survey modes, which are the modes most commonly used in business settings. This impact becomes even more pronounced when combined with a new statistical technique that has achieved great prominence within the business community – net scoring. Net scoring is most commonly associated with the Net Promoter Score, which Frederick Reichheld introduced in a *Harvard Business Review* article in 2003. The process for calculating a "net score" creates significant threshold effects[1]. Small changes in respondents' survey scores can have dramatic effects in the net scores.

Our research shows how the administration mode bias can greatly distort the so-called Net Promoter Scores *within* a company when mixed-mode surveying is conducted. By extrapolation, using NPS as a cross-company evaluative measure becomes highly problematic if the surveying practices are not identical, including the administrative mode. While the research was not conducted as a perfectly controlled experiment, common with research performed in real world settings, the findings are quite strong to show that telephone survey administration mode leads to higher survey scores than web-form survey administration mode, causing significant differences in means and in Net Promoter Scores.

Background – Survey Mode and NPS

Most of the research on survey mode bias lies in the public health, public opinion, political, and social science fields. These heavily focus on contrasts among face-to-face interviews, telephone surveys, and paper-based surveys. Research on the impact of web-form surveys versus other modes is new, and no studies appear to have been done on the impact of mode bias using data from the business domain. Yet, businesses are one of the major users of survey research.

Companies conducting customer feedback surveys are always challenged to get response rates up. (Nunley 2013) Some use mixed mode administration to get higher response rates, matching mode to the preferences of the respondents. However, the advantages of mixed mode administration may be outweighed by the biases introduced to the data set collected.

Bias occurs when the data collected do not accurately describe the true feelings of those being researched. Many different types of biases can be found in surveys. Some of these biases share common sources in survey practices, and in most surveys multiple biases will be in play. Survey respondents are not a homogeneous group, and various effects likely affect respondents differently. This makes it difficult to isolate the effects.

The sources of error in surveys may result from non-measurement errors, such as sample selection and non-response issues, and measurement errors, such as instrumentation and administration biases. (Bowling 2005) Mode bias may be reflected in several of these effects, but in general mode bias is where the mode creates a different mental frame for respondents (Groves 1999) affecting *whether* they respond and *how* they respond, presenting non-measurement and measurement errors, respectively.

Telephone surveys are favored by many surveyors, including in the business-to-business (B2B)

[1] NPS, and Net Promoter Score are trademarks of Satmetrix Systems, Inc., Bain & Company, and Fred Reichheld.

setting, because they reduce many *non-measurement errors*. *Non-response bias* is the impact on the sample statistics that results from invited respondents choosing to not respond where those non-respondents have perceptions that structurally differ from those who did choose to respond. (Bowling 2005) For example, those with strong opinions are more self-motivated to take a survey, so those lacking strong feelings are likely underrepresented in the sample data, causing a non-response bias. Peress (2010) and others have developed models to adjust for the non-respondents, but arguably the best approach is to get higher response rates.

The active solicitation of respondents through phone calls typically garners higher response rates, especially for those less motivated, and also higher completion rates for specific question items. (Groves 1989; Bredeson 2013; Nunley 2013) Thus the higher telephone response rates reduce the likelihood of non-response bias. The web-form survey process, in contrast, has respondent self-selection, lowering response rates, especially for the less motivated, and increasing the likelihood of non-response bias.

The choice of survey mode does create a *composition effect*, another non-measurement error. People with certain demographic profiles, such as age, are more or less likely to take a survey based upon the specific survey mode, biasing the survey results in a similar way that a sample selection bias would.[2] Some organizations use mixed-mode survey procedures to reduce this composition effect.

Survey mode also affects the type and degree of *measurement errors* for a survey. The measurement effect is where the survey mode affects how people respond to survey questions. (Bowling 2005; Voogt and Saris 2005; Weisberg 2005). Consider the measurement challenges inherent in any survey. For each survey question, respondents must comprehend the question, comprehend how they are being asked to respond, recall any relevant information that pertains to the question, develop an evaluation, and communicate a response through the survey medium. (Tourangeau 1984). Each survey mode places different burdens upon the respondent, and three factors differentiate these modes: technical factors specific to the mode, the role of the interviewer if any, and the communication medium for the survey mode. (de Leeuw 1992, 2005).

Telephone surveys present several challenges for controlling measurement bias. First, the cognitive demands made upon respondents are greater than for paper or web-form surveys, which could lead to responses that do not accurately reflect the respondents' views. For telephone surveys, oral transmission is typically the sole communication medium, and the use of verbal and numerical descriptors must be more succinct to reduce respondent burden. For this reason interval-rating questions are commonly used. Multiple survey questions can be posed using the same response scale, lowering respondent burden and helping survey completion rates.[3] (Nunley 2013) On web-form surveys, the question and the response scale are presented visually with some combination of verbal, numerical, and iconographic descriptors, simplifying the comprehension, evaluation, and response tasks for the respondent.

[2] Sample selection bias is where the surveyor chooses a sample that is a biased subset of the population.

[3] The assumption of interval properties for these scalar rating questions is questionable, but such questions can always be analyzed as ordinal data, as is done with "net scoring." In our experience few businesses are aware of the interval property requirement to legitimately calculate mean scores.

Second, the presence and actions of the interviewer may introduce measurement error. Just the presence of an interviewer can create a sense of time urgency for the respondent. Also, *interviewer bias* will be introduced if all the interviewers do not use the same intonation in presenting the question and provide the same guidance or "rebuttals" to queries from the respondent if they do not understand a question. In contrast, all respondents to web-form surveys receive the same stimuli increasing reliability in comparison to telephone surveys.

Third, past research has suggested that telephone surveys exhibit a *response effect* resulting from acquiescence, social desirability, and primary and recency effects. The presence of the interviewer, even absent any interviewer bias, is likely to solicit more positive responses to survey questions than a self-administered web-form survey. This is the *acquiescence effect*, also known as "yes saying." (Bowling 2005) Interval-rating questions using a strength-of-agreement scale are particularly prone to this effect. Respondents may also follow social norms in formulating their responses to look favorable to the interviewer, which is a *social desirability effect.* We suspect that in business-to-business surveys the social desirability effect is less of a factor in causing higher scores in telephone surveys since respondents are evaluating products and services for which they have paid.

Past research has shown that telephone survey mode received higher scores than survey modes with a visual presentation of the scale, for example, with a card displaying the scale in a face-to-face interview. (Groves 1979; Jordan *et al.* 1980). Dillman and Mason (1984) found that telephone surveys received higher scores than mail surveys, and Tarnai and Dillman (1992) found even providing a paper copy of the questionnaire for the telephone respondent still resulted in higher scores for telephone surveys over paper mail surveys. Dillman *et al.* (2001) found the same result for telephone respondents with a paper questionnaire versus web surveys.

More recent studies (Kreuter 2008; Bethlehem 2012; Christian *et al.* 2007) have found that responses were more positive for telephone surveys than for web-form surveys, including a tendency to select the most positive extreme endpoint category. Christian *et al.* also found for their college student sample that responses to telephone surveys garnered higher scores than web-form surveys with similar scale designs. However, they found that for telephone surveys, scales that were endpoint anchored received lower scores than scales with verbal descriptors for each point. Professional surveyors for businesses find they get higher scores from telephone surveys as well. (Bredson 2013; Nunley 2013)

Primary and recency effects result when respondents are drawn to the first or last response option presented to them. These effects are particular problematic for telephone surveys because of the manner in which interval-response scales are presented to respondents. In most all telephone surveys, regardless of the response scale length, interval-rating questions are presented as endpoint-anchored response scales. That is, the question is typically read by the interviewer to the respondent as, "On a scale from X to Y, where X represents <low anchor> and Y represents <high anchor>, how would you rate <construct>…"

This more succinct approach lowers the cognitive burden as opposed to presenting verbal anchors for each point on the scale; however, it increases the likelihood of the respondent choosing the first or last point on the response scale. For longer response scales, e.g., more than 7 response points, endpoint anchoring is typical for any survey mode, though on web-form surveys

verbal anchors are sometimes placed over a midpoint or over pairs of points on the scale. For telephone survey mode, endpoint anchoring is essential for longer scale lengths.

Even with verbal anchors for each point on the response scale, some research has shown that telephone surveys tend to get more extreme responses than face-to-face interviews. (Nicholaas *et al.* 2000; Sudman *et al.* 1996) While generally thought that the recency effect would trump the primary effect, Christian *et al.* (2007) presented the scales in alternate order and without either primacy or recency dominant.

Due to the manner in which response scales are presented to the respondent, we suggest that the observed primary and recency bias may be more accurately described as a *response scale truncation effect*. The respondent is asked to consider where their views fall along the range of the scale, but all they hear are the endpoint anchors. The time pressure to provide a response means they are less likely to consider points on the scale other than the endpoints.

In essence, the scale becomes a binary scale, composed of the low and high anchors. Telephone respondents would be more likely to select what they hear, which are the scale endpoints. Thus the response scale, regardless of its actual length is truncated. If respondents have any positive view of the phenomenon being measured by the survey question or if an acquiescence effect occurs, then they are more likely to select the top end response option, which was orally presented to them. This truncation effect would lead to more extremes in the distribution of scores and in the calculated statistics, especially the net scoring statistic.

The Net Promoter Score (NPS) is a customer feedback approach developed and promoted by Frederick Reichheld and Satmetrix, a surveying vendor on whose Board he sits. In his *Harvard Business Review* article in 2003, "The One Number You Need to Grow," Reichheld outlines a multi-phase research stream in which he identified the "likelihood to recommend" survey question as the single best indictor of the future profitability of a firm. (Reichheld, 2003) The recommendation question had been asked on customer surveys as an attitudinal indicator for many decades, along with survey questions asking "overall satisfaction" and "likelihood of future purchase." (Bredeson 2013; Nunley 2013; Tarter, 2013) However, Reichheld's study found predictive power not previously expected for the recommendation question.

Reichheld's stated goal was to make the results from the recommendation question more actionable, that is, to identify at-risk customers and drive front line management to address the concerns of these customers. To help accomplish this, Reichheld added a new analytical approach to the recommendation survey question. He created a "net scoring" statistic. Net scoring, like the mean, is a single statistic to summarize a data set. The logic of net scoring is to take the percentage of respondents at the top end of the response scale, so-called "top box," and subtract from it the percentage of respondents at the lower end of the response scale, so-called "bottom box." Net scoring thus arrives at a single number, which is expressed as a percentage that can range from +100% to -100%. While this statistic can be calculated for any ordinal or interval-rating survey question, Reichheld applied it to the recommendation question because of the predictive power he found for that question.

The idea of so-called "top box" and "bottom box" scoring had been practiced in the customer surveying industry since the 1980s as a method to present survey results, and specifically the dispersion of responses, in a manner more understandable to managers than means and standard

deviations. (Nunley 2013; Tarter, 2013) In particular, the "bottom box" draws attention to the responses at the lower end of the response scale, which can become masked when viewing only mean scores. Managerially, the low end is important as these likely represent at-risk customers. Also, the easiest way to raise the mean score is to improve the responses given by those scoring at the low end as there is more opportunity for greater improvement.

Operationally, Reichheld chose to use an 11-point scale for the recommendation question ranging from 0 to 10. He argued that the longer scale provided more precision and that having a zero on the scale clearly defined the direction of the scale to reduce the likelihood of scale-inversion scoring errors by the respondent. In practice, we see many companies using this 11-point scale just for the recommendation question while using a different length scale for other questions in the survey.

He defined the "top box" as those providing scores of 9 or 10, which he labeled as "Promoters," and the "bottom box" as those providing scores of 0 to 6, which he labeled as "Detractors." Those providing 7s and 8s were labeled as "Neutral" or "Passives." Thus the net score is (9s+10s) – (0s to 6s), expressed as a percentage. The idea of subtracting the bottom box from the top box was presented by Sambandam and Hausser in a 1998 online paper; however, they provided a double weighting for the bottom box score to give it more impact upon the resulting percentage. Net scoring did not become an industry practice until Reichheld's article in 2003.

It was this metric, net scoring of the recommendation question, which he curiously labeled Net *Promoter* Score, that Reichheld's study found to have strong predictive value for future profitability. While some researchers have found support for Reichheld's contention (Marsden *et al.* 2005), his findings have been challenged by academics. (Morgan and Rego 2004; Keiningham *et al.* 2007; Keiningham *et al.* 2008) Many practitioners have identified issues with the findings and in particular how NPS has been applied at companies. (Plowman; Grisaffe)

NPS has taken hold as key customer metric as evidenced by the many industry conferences and LinkedIn discussion groups dedicated to this metric. NPS and Net Promoter have entered the lexicon, including as part of job titles. Companies are no longer doing "relationship surveys," they are conducting "Net Promoter surveys." While Reichheld viewed NPS as a way to drive accountability at the front lines by having front line managers engage in a timely manner those customers who provided low scores, NPS has become in practice a summary performance metric. Further, NPS has become viewed as an "industry metric."[4]

Serious validity issues are present when comparing survey scores across companies where the survey instrument and the survey administration practices are not standardized. Yet, the authors are familiar with many companies that look to benchmark scores from surveys they conduct against published Net Promoter Scores available on the internet without any consideration of the shortcomings of the comparison, including the impact of survey mode. Our research shows the impact of survey mode upon survey scores, and it shows the impact that survey mode can have upon the net scoring statistic due to the threshold effects inherent in its calculation.

[4] See for example: http://tinyurl.com/btkmmdq

The Research Setting

Our research hypothesis is that the differences in survey scores in the data set we analyzed result from the response effect in mode bias. Additionally, the compositional effect would impact the summary statistics when the mode composition changes in mixed mode surveys.

The company whose data we analyzed wished to remain anonymous, so we will be referred to as Pictor. This large company provides sophisticated products to business users, not to end consumers, that is, it's a so-called B2B company, and it provides services to maintain this equipment on a contractual basis. Accordingly, Pictor has a large field organization, organized into 133 districts, whose personnel interact with end users on a regular basis. Each year the company has about 750,000 transactions with customers. The majority of the interactions involve performing preventative maintenance and calibrations on installed equipment.

The company conducts transactional surveys to assess the customers' experience with their most recent service as well as measuring their overall satisfaction and loyalty using the Net Promoter question. Every two weeks, a list is randomly generated of 5200 customers with whom Pictor has had a service transaction the previous fortnight. Anyone who has been surveyed in the past six months is removed from the list to limit survey fatigue.

How each customer is surveyed is based upon whether the company has an email address for the customer. Customers for whom Pictor has an email address receive an email invitation to take the survey, which contains a link to a web-form survey. One week after the first invitation, a reminder email will be sent to those who have not responded. If the company does not have an email address for the customer, then the surveying is done by telephone interview. Six attempts will be made to contact each invitee by telephone. Both the web-form and telephone surveys are conducted by an independent, third-party professional surveying organization.

The survey instrument language is identical regardless of which method is used for conducting the survey. It consisted of the following three questions.

- Assuming you were allowed to do so, how likely would you be to recommend Pictor to colleagues within your organization or to other organizations?
- Please rate your overall satisfaction with Pictor as a [product type] service provider.
- Overall, how satisfied were you with this most recent service visit?

All questions are posed on a 0-to-10 scale, though our analysis does not assumes interval properties for the data. The telephone surveys are presented with endpoint verbal descriptors only. The web-form survey has endpoint verbal anchors with numerical descriptors for each response point. Note that Pictor has modified the "standard" Net Promoter question with the qualifying clause, "Assuming you were allowed to do so." One criticism of NPS for B2B companies is that employees may be prohibited from making recommendations. Thus, Pictor added this phrasing to increase the ability of people to provide a response, yet many respondents still declined to answer due to their company's prohibition. In the web-form survey, respondents were not required to answer every question.

The survey data are all imported into a single Enterprise Feedback Management (EFM) system for analysis. Pictor does not examine the results separately by survey mode, i.e. phone or email. The company uses the results to drive action by field managers, but they are also used as performance measurements of management.

We were provided access to data for December 2011 for surveys conducted in the United States and Canada. Thus, we do not have cultural effects in the survey data measurement that would be introduced by surveys conducted worldwide. For that month, Table 1 presents basic survey statistics by survey mode. Response rate is not tracked by survey mode at Pictor. As noted, some phone respondents indicated they could not answer the recommendation question, even with the qualifying clause mentioned above, or that the question was not applicable. For the web-form surveys, responses were not required, leading to item non-response. These response records were removed the data set. The difference between the Number of Responses and the Usable Responses represents those records removed. Note that because web-form surveys are self-administered, the Percent Unusable Responses is far higher than for telephone surveys where the interviewer is more likely to push for a response and the respondent will feel more compelled to provide a response. The effect of this item non-response upon the summary data cannot be measured, but it could account for some of the discrepancy between survey modes.

Table 1
Summary Survey Statistics
December 2011

	Overall	Telephone	Web-Form
Number of Survey Invitations	4718		
Number of Responses	3079	2855	224
Response Rate	65.30%		
Usable Responses		2711	178
Percent Unusable Responses		5.0%	20.5%
Usable Responses from Districts with Web-Form Responses		1946	178

For this to have been a perfect controlled experiment, customers would have been chosen at random to be solicited for the survey either by telephone or by email invitation. That was not the case here. The company has been slowly adding email addresses for its customers, but it is the responsibility of the local field office to add this information to the customer record.

Of the 133 districts in the US and Canada, 48 districts had no email responses at all. Those 48 districts represented 30.2% of the telephone surveys completed. To eliminate one potential confounding variable in the analysis of survey mode that follows, we only used data for those districts that had some email responses, reasoning that the districts that have not bothered to collect email addresses for its customers might be uniquely different. This left us 1946 telephone surveys and 178 web-form surveys as shown in the bottom row of Table 1.

We did test to see if there was a difference in the telephone scores between districts that do and do not collect email addresses. A chi-square test was performed on the Recommendation question only for the telephone survey responses. Due to low expected values, counts for the 0 and 1 points on the response scale were combined as well as counts for the 2 and 3 points on the response scale to get expected values of at least 5 as required for the chi-square test. The p-value was 0.65, indicating no statistically significant difference between the two groups. While not a statistically significant difference, scores were slightly higher for districts with no email addresses. Therefore, including the 48 districts with no email addresses would have made the differences between survey modes even more extreme than shown below.

Regardless, the threat to validity still exists that the customers from whom the company has collected their email addresses are structurally different from customers for whom the company does not have an email address. We can only speculate about whether this form of a selection effect – a surveyor selection effect as opposed to a respondent self-selection effect – is a proxy for some other bias. No clear answers could be given about why or why not email addresses had been collected by field offices. Those with email addresses in the database might be newer customers or customers with whom there has been more recent service contact.

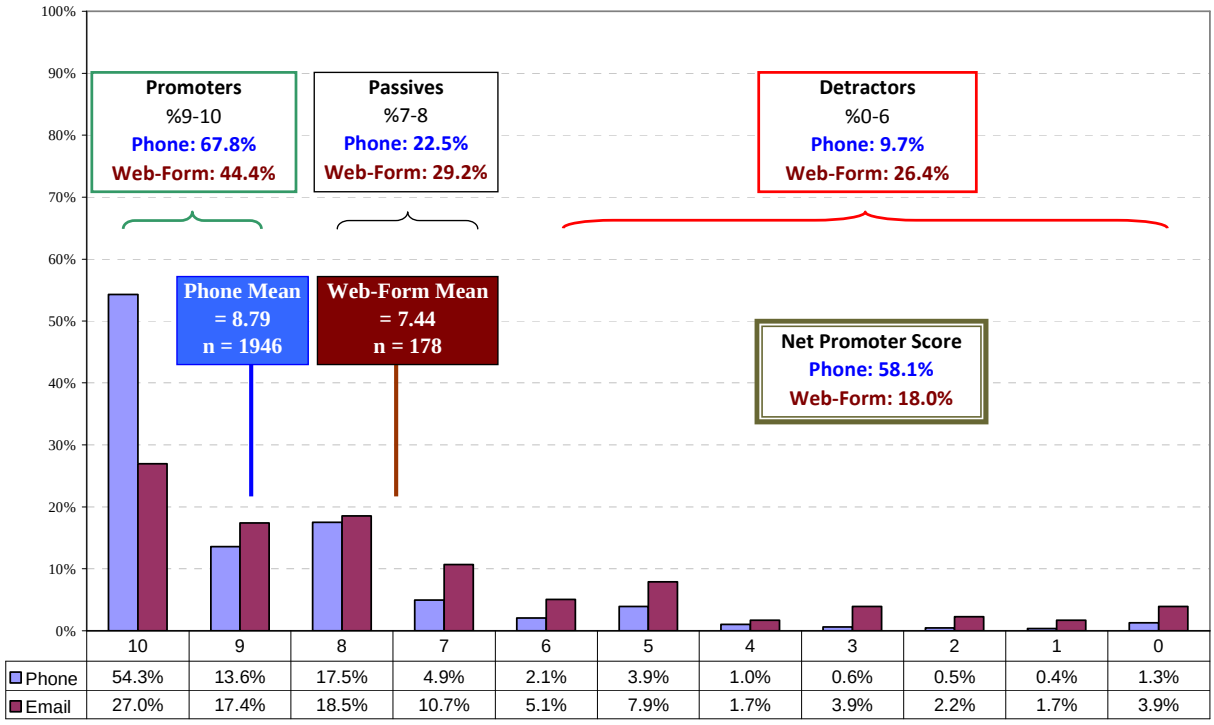
As will be shown, the differences in the scores by survey mode are so distinct it is unlikely that this factor alone could explain those differences. The reality is that few companies are willing to conduct true experiments, so even this compromised experiment is valuable.

Data Analysis and Findings

To illustrate the impact of survey administration mode upon survey results, we analyzed the month's worth of data from Pictor. We examined all three questions in the survey, listed previously. Our focus, however, will be on the so-called Net Promoter question.

Figure 1 shows the frequency distribution of survey responses for the Recommendation question, broken out by survey method. The differences in the distributions are dramatic, particularly at the top end of the distribution. While 54% of phone respondents gave scores of 10, "only" 27% of email respondents gave 10s, a 2:1 ratio.

Figure 1
Frequency Distribution Phone vs. Web-Form
Recommendation Question with Net Promoter Scoring
December 2011 Data



	10	9	8	7	6	5	4	3	2	1	0
Phone	54.3%	13.6%	17.5%	4.9%	2.1%	3.9%	1.0%	0.6%	0.5%	0.4%	1.3%
Email	27.0%	17.4%	18.5%	10.7%	5.1%	7.9%	1.7%	3.9%	2.2%	1.7%	3.9%

For every other point on the scale, those who responded via web-form had higher frequencies. In the bottom half of the distribution, almost no phone respondents gave scores while there were some scores given here by web-form respondents. Simply put, the phone survey method garnered far more scores in the top response option, and the web-form surveys had a more even distribution. As shown on the chart, the difference between survey methods was also pronounced in the mean score for each: a mean of 8.79 for phone surveys versus a mean of 7.44 for the web-form survey.[5]

To determine whether the difference seen between the two survey modes was statistically significant, we conducted a chi-square test. When using separate categories for each of the 11 points on the response scale, the expected values for the scale points at the lower end were less than the required 5. Accordingly and after experimenting with several combinations, we grouped scale points 1 to 4 together and scale points 5 and 6 together. This yielded us the distribution shown in Figure 2. The chi-square test statistic was 81 versus a critical value of 11.1, resulting in a p-value of 5.09 E-16. The difference between survey modes was highly significant. As visually implied by the chart, the largest chi-square values were for the 10 and 0-to-4 categories where

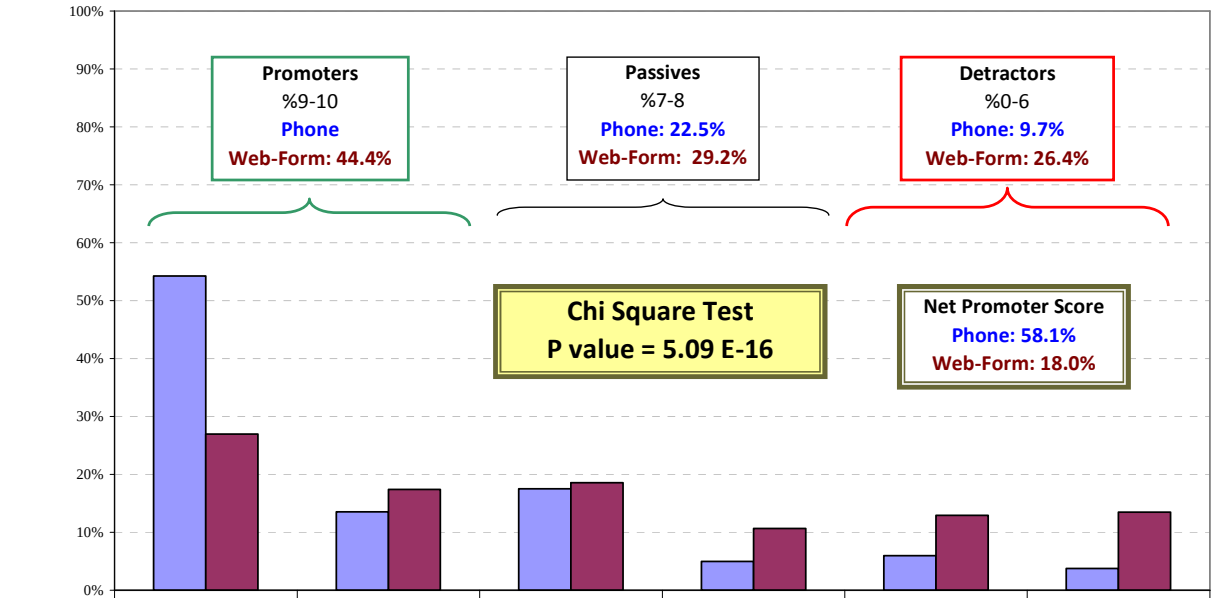
[5] While an assumption of interval properties is doubtful for these data, we did conduct a t-test assuming unequal variances. (The F-test showed that the variances between the two data sets were unequal. The p-value for a two-tail test was 2.02 E-12) The t-test showed that the difference in the mean scores was not due to random sampling error. The p-value for a two-tail test was 8.13 E-10.

the difference in survey scores was the greatest.

To be certain that the difference in survey mode was not an effect caused by something related to the recommendation question, we also conducted chi-square tests for the other two questions in the survey. For the overall satisfaction question, the p-value was 2.29 E-16, and the response pattern across the categories was similar to the recommendation question. For the visit satisfaction, the differences between survey modes was less dramatic; however, the differences were still statistically significant with a p-value was 0.0035.

We have noted that this is not a perfect experiment, and it is possible that the effects shown here are not due to mode but do to a bias introduced by how Pictor collects email addresses from its customers. However, the high levels of statistical significance make it unlikely that this confounding effect could explain the differences observed.

Figure 2
Frequency Distribution & Chi Square Test Results
Phone vs. Web-Form for Recommendation Question
December 2011 Data



	10	9	8	7	5 to 6	0 to 4
Phone	54.3%	13.6%	17.5%	4.9%	6.0%	3.8%
Web Fom	27.0%	17.4%	18.5%	10.7%	12.9%	13.5%

As stated earlier, many practitioners choose to use cumulative frequency summary statistics to describe the message in the survey data. Aside from eliminating the assumption of interval properties needed for the mean statistic, practitioners believe that top box, bottom box, and net statistics provide focus to the lower end of survey scores to help drive operational implementation of the findings of the survey. Here, we see the impact of survey mode upon these statistics.

Table 2 shows the data, also depicted in Figures 1 and 2, of these alternative statistics practitioners use to summarize survey scores. The data in Table 2 show how this shift in the

distribution of scores between modes is amplified by the Net Scoring logic. A difference in the mean score from 8.79 to 7.44 is certainly enough to pique most any manager's interest, but a difference in the net scoring for the recommendation question (NPS) from 58.1% to 18.0% is of crisis-level proportions.

Table 2
Top Box, Bottom Box and Net Promoter Scores for Recommendation Question
By Survey Mode

Survey Method	Top Box Promoters (9s+10s)	Passives (7s+8s)	Bottom Box Detractors (0s to 6s)	Net Score (Promoters – Detractors)
Phone	67.8%	22.5%	9.7%	58.1%
Web-Form	44.4%	29.2%	26.4%	18.0%
Combined Phone & Web	65.8%	23.1%	11.1%	54.7%
Combined with 10% shift to Web	63.5%	23.7%	12.8%	50.7%

While the mode bias has led to different distributions, the dramatic difference in the cumulative frequency statistics is due to the threshold effects inherent in these statistics. A shift in respondents scoring from 10s to 9s, 8s to 7s, or 6s to 0s would have no impact upon the scores in Table 2, but shifts from 9s to 8s or 7s to 6s (or the reverse) causes changes in the statistics. In this research, the shifts in scoring were due to changes in survey mode, meaning the difference in Net Promoter Scores between modes is measurement error and not actually reflective of changes in the perceptions of the respondents.

To give a sense of the implications for mixed-mode surveying, we looked at what would happen with a 10% shift in surveying from phone to web-form mode. We have 2124 total surveys in the data set. (See Table 1.) What if 10% of those (212) shifted from telephone survey mode to web-form mode? Currently, the NPS is 54.7% for the two modes combined, which is how Pictor analyzes its data. (See Table 2.) With a 10% shift in modes used and the scoring by mode following current trends, the combined NPS would drop from 54.7% to 50.7%. Thus, as Pictor gathers more email addresses for its clients and the mode mix shifts, we would expect the NPS to drop, *ceteris paribus.*

We also analyzed the data for two other questions in the survey for mode effects and found results for the overall satisfaction question similar to the recommendation question; the net scores were 65.9% for phone and 29.4% for web-form. For the visit satisfaction question, we saw overall more positive scores, especially from the web-form respondents. The net visit satisfaction scores were 73.7 % for phone and 56.9% for web-form. Recall that the recommendation and overall satisfaction questions were positioned for the relationship overall; whereas, the visit satisfaction question pertained to the last service visit.

Implications

Given the strong evidence that survey mode matters, the real question is how this should be handled by organizations that collect feedback from various stakeholders using mixed-mode data. Extraneous or confounding factors must be adjusted for or eliminated for survey findings to be credible and meaningful. A good solution must be accurate but also acceptable to management teams in organizations.

For Pictor, they face an interesting analytical dilemma. As shown, as a greater percentage of surveying is done by web form, the summary statistics will decline purely due to the survey mode effect. Thus, it behooves organizations such as Pictor to give strong thought to their surveying practices and adopt one of the following approaches.

Develop adjustment factors. One option is to develop algorithms for adjusting scores from multiple modes to present consistent data across modes. This work is in progress in select areas (Vannieuwebhuyze 2010; Elliott *et al.* 2009; Schonlau 2006), but it seems unlikely that one set of adjustment factors can be applied across any survey for any organization. The adjustment factors would likely need to be developed for each unique situation. In practice, that is unlikely to happen on any large scale due to the costs, especially for smaller organizations. Additionally, explaining to management and gaining acceptance for the adjustments would be a challenge. This would be especially true where the survey findings are used for performance evaluations. When survey results are poor, the first inclination is to challenge the methodology, and the adjustment factors would be one more element to be challenged.

Change survey delivery practices. As Christian *et al.* found, how the questions are presented by telephone does make a difference to reduce the scale truncation effect that is likely in play. This requires fully anchored scales for telephone survey mode, which are higher in administrative and respondent burden. For an 11-point scale used for the so-called net promoter question, fully anchored scales are simply impractical for telephone administration.

Track mixed-mode administrations separately. If mixed mode is deemed necessary due to the contact information for the research population or the preferences of the research population, then tracking and trending the data separately by administrative mode would be simplest to explain to management. This practice would also highlight the differences by mode; however, it would make it more difficult to create a summary score and make comparisons across organizational units.

Discontinue mixed-mode survey administration. Perhaps the simplest way of addressing the confounding effects from mixed-mode surveying is to stop doing it. If an organization is using a survey program purely for internal purposes, especially trend analysis, then consistency across survey administrations is very much within its control.

Educate consumers of survey data. Perhaps the most important implication from this research is the need to educate the business community on the shortcomings of survey data. In particular, companies that use mixed-mode surveying need to be aware of how changes in the mix can dramatically change survey results.

Also, companies should be very circumspect when comparing data between two totally different

surveys that may share questions that purport to measure the same underlying construct. Without knowledge of all the surveying practices – both the survey instrument design and the administration practices – the differences seen between survey data may have little to do with true differences in the perceptions of stakeholders being surveyed. The differences may in fact be artifacts of the surveying practices. In regards to net scoring, consumers of such statistics need to understand exactly what the statistic implies and the threshold effects that the calculation creates. Differences in survey practices can be amplified by the net scoring logic, leading to incorrect business decisions.

References

Bethlehem,, J.; Biffignandi, S. *Handbook of Web Surveys*. Wiley Handbooks in Survey Methodology 567. New Jersey: John Wiley & Sons, 2012.

Bowling, Ann, "Mode of questionnaire administration can have serious effects on data quality," *Journal of Public Health,* 2005, Vol. 27, No. 3, pp. 281–291.

Bredeson, Jean Mork, President, Service 800, interview conducted April 1, 2013.

Christian, Leah Melani, Don A. Dillman, and Jolene D. Smyth, "The Effects of Mode and Format on Answers to Scalar Questions in Telephone and Web Surveys." in J. Lepkowski, C. Tucker, M. Brick, E. DeLeeuw, L. Japec, P. Lavrakas, M. Link, and R. Sangster (Eds.) *Advances in Telephone Survey Methodology*. New York: Wiley- Interscience, 2007.

de Leeuw, Edith D., *Data Quality in Mail, Telephone, and Face to Face Surveys*. Amsterdam: TT Publications, 1992.

de Leeuw, Edith D., "To Mix or Not to Mix: Data Collection Modes in Surveys." *Journal of Official Statistics*, 2005.

Elliott, Marc N., Alan M. Zaslavsky, Elizabeth Goldstein, William Lehrman, Katrin Hambarsoomians, Megan K. Beckett, and Laura Giordano, "Effects of Survey Mode, Patient Mix, and Nonresponse on CAHPS® Hospital Survey Scores," *HSR: Health Services Research* 44:2, Part I, April 2009, pp. 501-518.

Frauke Kreuter, Stanley Presser, and Roger Tourangeau. "Social Desirability Bias in CATI, IVR, and Web Surveys: The Effects of Mode and Question Sensitivity". *Public Opinion Quarterly*, 2008, vol. 72, No. 5, pp. 847-865

Grisaffe, Doug, http://www.xzamcorp.com/resource-center/48-nps/110-net-promoter-score-problems.html, no publication date.

Groves, Robert M., *Survey Errors and Survey Costs*, New York: Wiley-Interscience, 1989.

Keiningham, Timothy L. *et al.*, "A Longitudinal Examination of Net Promoter and Firm Revenue Growth," *Journal of Marketing,* Vol. 71, July 2007, pp. 39–51.

Keiningham, Timothy L. *et al.*, "Linking Customer Loyalty to Growth," *Sloan Management*

Review, Summer 2008, pp. 51-57.

Marsden, Paul, *et al.*, "Advocacy Drives Growth: Customer Advocacy Drives UK Business Growth," *Brand Strategy,* Nov.-Dec. 2005.

Morgan, Neil A. and Lopo Leotte Rego, "The Value of Different Customer Satisfaction and Loyalty Metrics in Predicting Business Performance," *Marketing Science,* Vol. 25, No. 5, September–October 2006, pp. 426–439.

Michael Peress, "Correcting for Survey Nonresponse Using Variable Response Propensity," *Journal of the American Statistical Association,* Volume 105, Issue 492, 2010.

Nicholaas, Gerry, Katrina Thomson, and Peter Lynn. *The Feasibility Of Conducting Electoral Surveys In The UK By Telephone.* Centre for Research into Elections and Social Trends. London: National Centre for Social Research, and Department of Sociology, University of Oxford, 2000.

Nunley, Roger, Managing Director, Customer Care Institute, interview conducted March 15, 2013.

Plowman, Howard, "Net Promoter Score – The Search for the Magic Pill," www.infoquestcrm.co.uk/Net-Promoter-Score.pdf, no date of publication.

Reichheld, Frederick, "The One Number You Need to Grow," *Harvard Business Review,* Nov-Dec., 2003, pp. 46-54.

Sambandam, Rajan and George Hausser, "An alternative method of reporting customer satisfaction scores," *Quirks Marketing Research Media,* 1998.

Sudman, Seymour, Norman M. Bradburn, Norbert Schwarz, *Thinking about answers: the application of cognitive processes to survey methodology*. San Francisco: Jossey-Bass, 1996.

Tarter, Jeffrey, CEO Emeritus, Association of Support Professionals, interview conducted March 1, 2013.

Tourangeau Roger, "Cognitive sciences and survey methods," in Jabine T, Straf M, Tanur J, Tourangeau R, eds. *Cognitive aspects of survey methodology: building a bridge between disciplines.* Washington DC: National Academy Press, 1984.

Vannieuwenhuyze, Jorre, *et al.*, A Method For Evaluating Mode Effects In Mixed-Mode Surveys, *Public Opinion Quarterly,* Vol. 74, No. 5, 2010, pp. 1027–1045.