Published in final edited form as:

Public Opin Q. 2015 ; 79(2): 420–442.

AN EVALUATION OF PRIMARY DATA-COLLECTION MODES IN AN ADDRESS-BASED SAMPLING DESIGN

ASHLEY AMAYA[*],

Graduate student in the Joint Program in Survey Methodology at the University of Maryland, College Park, MD, USA, and formerly a survey methodologist in the Public Health Research Department at NORC at the University of Chicago, Chicago, IL, USA

FELICIA LECLERE,

Senior fellow in the Health Care Department at NORC at the University of Chicago, Chicago, IL, USA

KARI CARRIS, and

Vice president of the Health Sciences Department at NORC at the University of Chicago, Chicago, IL, USA

YOULIAN LIAO

Epidemiologist at the Centers for Disease Control and Prevention, Atlanta, GA, USA

Abstract

As address-based sampling becomes increasingly popular for multimode surveys, researchers continue to refine data-collection best practices. While much work has been conducted to improve efficiency within a given mode, additional research is needed on how multimode designs can be optimized across modes. Previous research has not evaluated the consequences of mode sequencing on multimode mail and phone surveys, nor has significant research been conducted to evaluate mode sequencing on a variety of indicators beyond response rates. We conducted an experiment within the Racial and Ethnic Approaches to Community Health across the U.S. Risk Factor Survey (REACH U.S.) to evaluate two multimode case-flow designs: (1) phone followed by mail (phone-first) and (2) mail followed by phone (mail-first). We compared response rates, cost, timeliness, and data quality to identify differences across case-flow design. Because surveys often differ on the rarity of the target population, we also examined whether changes in the eligibility rate altered the choice of optimal case flow. Our results suggested that, on most metrics, the mail-first design was superior to the phone-first design. Compared with phone-first, mail-first achieved a higher yield rate at a lower cost with equivalent data quality. While the phone-first design initially achieved more interviews compared to the mail-first design, over time the mail-first design surpassed it and obtained the greatest number of interviews.

[*]Address correspondence to: Ashley Amaya, Joint Program in Survey Methodology, 1218 LeFrak Hall, University of Maryland, College Park, MD 20742, USA; aamaya@umd.edu.

The findings and conclusions in this paper are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention or NORC at the University of Chicago.

Introduction

Address-based sampling (ABS) via the US Postal Service computerized Delivery Sequence File (DSF) has emerged in the past decade as the sampling frame of choice for a wide variety of surveys. First adopted as a cost-saving alternative to field listing (Iannacchione, Staab, and Redden 2003; O'Muircheartaigh, Eckman, and Weiss 2003; Montaquila, Hsu, and Brick 2011), ABS using the DSF has been adopted only recently as an alternative to random-digit dialing (RDD) (Iannacchione 2011). The DSF provides coverage of nearly all US households (Montaquila et al. 2009; Fahimi 2010)—including those segments of the population known as cell-phone-only users and those without telephone service. Others have found the ABS frame useful to conduct effective mail surveys with response rates comparable or superior to a traditional RDD frame (Link, Battaglia, et al., 2008; Brick, Williams, and Montaquila 2011).

The advent of ABS designs and the promise of nearly complete coverage of households have been accompanied by growing interest in and resurgence of mail surveys and a variety of data-collection efforts (Couper 2010; Groves 2011). ABS affords researchers great flexibility in the choice of the initial data-collection mode and the sequencing of modes. With an address, researchers can mail self-administered questionnaires (SAQs) or web survey access instructions to sampled respondents, attempt to contact respondents by telephone (provided that a phone number can be reverse-matched to the address), or visit the sampled address to conduct an in-person interview. Indeed, researchers surmise that an ABS design coupled with multiple data-collection modes has great potential for reversing declining survey response rates (de Leeuw 2005; Groves 2011) and improving population coverage while controlling costs (Iannacchione, Staab, and Redden 2003; Link, Daily, et al. 2008; Link, Battaglia et al. 2008; Williams et al. 2010; Brick, Williams, and Montaquila 2011).

Given the promise of ABS multimode designs, a growing body of research has been conducted to inform survey best practices. Previous research on the operational aspects of ABS multimode designs has focused on efficiencies within a particular multimode design (e.g., methods to screen households in a mail and telephone design (Murphy, Harter, and Xia 2010) and comparing response rates of multimode designs to single-mode designs (Messer and Dillman 2011). Additional work also has been conducted on the sequencing of modes, but this research has been limited to the ordering of web and mail multimode designs (Messer and Dillman 2011; Millar and Dillman 2011).[1]

Missing from the research is a clear comparative framework in which to choose a starting mode for data collection when considering telephone and mail. Moreover, the previous literature has focused almost entirely on response rates and has rarely considered other measures that may inform best practices. In this paper, we report the results of an experiment that we conducted to evaluate two ABS multimode case flows: phone followed

[1]While this work does not use the Delivery Sequence File as the sampling frame, it still uses an address-based frame and is worth including here given the experimental multimode design.

by mail (“phone-first”) and mail followed by phone (“mail-first”). We use response rates, cost, timeliness, and data quality to assess the efficiency of each case-flow design.

Background to the Problem

ABS multimode designs require researchers to make decisions about initial data-collection modes and the choreography of subsequent modes. In recent years, methodologists have tested several approaches for combining multiple data-collection modes within the context of an ABS design. To date, the operational focus has been threefold: (1) research to improve the efficiency of data collection given a particular sequence of modes; (2) research on whether multimode surveys are superior to single-mode surveys; and (3) the order of data-collection modes. Within the first avenue of research, individuals have investigated efficiency improvements for several combinations of modes—phone-first, mail-plus-web (in both mode orders), and mail-first. For example, Amaya, Skalland, and Wooten (2010) and Murphy, Harter, and Xia (2010) have investigated ways to prioritize telephone contacts where possible and to offer mail as a secondary mode—an efficiency improvement within a “phone-first” design. Other researchers have tested the effectiveness of a host of operational procedures on mail-plus-web surveys. Messer and Dillman (2011) found that the use of prepaid incentives improved the response rate and the proportion of individuals that participated via the web. In the same survey, the use of web instructional inserts and alternative postage methods did not affect overall response rates. Finally, efficiency research has been conducted on mail-first designs to test the effectiveness of varying incentive amounts and survey length on response rates. Montaquila et al. (2013) found that offering a $5 incentive significantly increased the response rate over the non-incentivized group, while an engaging (i.e., longer) screening interview improved response rates when the name of an individual was requested.

Researchers have also focused on the worthwhileness of multimode surveys compared to single-mode surveys. Brick, Williams, and Montaquila’s (2011) study analyzed the difference in response rate gains between nonresponders that were mailed additional SAQs to nonresponders that were transitioned to telephone contacts. Within a “mail-first” framework, no comparative advantage to mode switching emerged, thus suggesting that a single-mode approach that begins and ends in mail may be the most efficient data-collection strategy for an ABS design even where a screening interview is necessary. Similarly, Messer and Dillman (2011) concluded from a series of mail-plus-web experiments that a single consistent mode was the most successful design for ABS, despite previous research into the value of multimode designs in improving response rates. Multimode designs may encompass not only the mode in which the questionnaire is delivered but also the way in which the respondent is contacted. In a mail survey, Brick et al. (2013) experimented with the mode of contact, varying the postage method (first class versus priority) and mode of reminder/thank-you contact (outbound interactive voice response (IVR) versus postcard), and found that varying the form of contact between the phases (i.e., introducing outbound IVR between mailings) improved response rates.

While research on operational efficiency and multimode superiority has spanned a variety of modes, work into mode order has been limited to mail and web. Moreover, the work on

mode sequencing focuses primarily on response rates and does not evaluate it by other criteria. For example, Millar and Dillman (2011) reported that starting with a web mode and following up with nonresponders via mail achieved superior response rates when compared to beginning with a mail mode and following up with the web. Finally, previous work has done little to evaluate mode order by different populations. While not all work on mode sequencing has focused on the general population (e.g., Millar and Dillman 2011; Montaquila et al. 2013), a comparative analysis of mode sequencing across survey populations has not been conducted for any combination of data-collection modes.

In this paper, we take the next step in assessing initial mode assignment and sequencing in ABS designs by directly comparing the mode sequence of telephone and mail on a variety of criteria. We report the results of an experiment that systematically varied whether we initially contacted and attempted to interview sampled addresses by telephone or via mailed questionnaires. Specifically, we examine response rates, cost, timeliness, and the potential for mode effects in the multimode follow-up. We also assess whether the efficacy of initial mode assignment varies by survey eligibility rates given that some subpopulations may be harder to reach by mail and telephone.

Decisions about starting mode are currently made largely on the availability of contact information and perceptions of total cost. It is still unclear from the research which mode sequence is likely to allow the touted advantages of an ABS design to be realized. We conducted an experiment to assess the efficiency, population coverage, and measurement consequences of a design that begins either by mail or on the telephone from an ABS sampling frame, matched by a vendor to phone numbers. We hope to answer the question of which mode is an optimal starting point.

Data and Methods

To understand the impact of initial data-collection mode on survey performance and data quality, NORC at the University of Chicago conducted an experiment during year four (November 2011–September 2012) of the annual Racial and Ethnic Approaches to Community Health across the U.S. (REACH U.S.) Risk Factor Survey. REACH U.S. was a Centers for Disease Control and Prevention (CDC)–sponsored survey conducted in 28 communities across the country. We limit this description of the REACH U.S. design to the six communities in which we conducted our experiment. Each REACH U.S. community varied in size from an entire state to a small cluster of Census tracts. The questionnaire was the same across communities, but different subpopulations were targeted within each community. Each REACH U.S. community selected individuals from one or more of the following racial/ethnic groups: Hispanic, African American, or Native American. Eligibility rates varied significantly across communities, which provided us with a natural experiment in which to test the impact of the starting mode on different types of survey populations.

An ABS design was used in all communities, but sampling strategies varied slightly by community. We drew a simple random sample of addresses from the DSF in three communities. In the remaining three communities, we drew a stratified random sample by race/ethnicity. Additional information from InfoUSA was available for some addresses

regarding the race/ethnicity of individuals at a frame address. These addresses were placed in a high-density stratum, while the remaining addresses on the frame were placed in a low-density stratum.

To determine the eligibility of household members, a screening interview was conducted via mail or phone immediately preceding the topical interview. A person had to be 18 years or older, live or stay at the sampled address, and self-identify with the targeted racial/ethnic group(s) for that community to be eligible for the main interview. Up to two eligible persons per household were randomly selected for the topical interview. These selected household members were asked to complete a standardized health questionnaire (modeled after the Behavioral Risk Factor Surveillance System [BRFSS]) about health conditions, health behaviors, and preventive health care. The mail screener and topical interview were part of the same mailing.

Experimental Design

In year four, we launched a case-flow experiment to evaluate a phone-first design compared to a mail-first design. The experiment was fielded in six of the 28 REACH U.S. communities. Three communities were in Chicago and targeted African Americans and Hispanics. The other three communities included the entire state of Oklahoma (Native Americans); Richmond, Virginia (African Americans); and Philadelphia, Pennsylvania (African Americans). These six communities were chosen for two reasons. First, they represented communities with very different population sizes and eligibility rates. Variability in eligibility rates allowed us to test whether case-flow designs were sensitive to different eligibility thresholds. Second, these communities consisted of three areas that targeted predominantly English-speaking populations (Oklahoma, Richmond, and Philadelphia) and three communities that targeted English- and Spanish-speaking populations (Chicago). Materials were mailed in English and Spanish in the Chicago communities, and bilingual telephone interviewers were available when needed.

A total of 23,613 sampled addresses across the six communities were randomly assigned to a phone-first or mail-first condition (figure 1). An attempt was then made to match each sampled address to a telephone number; only cases that matched to a telephone number were retained for the experiment (n = 9,489, or 40.2 percent). The unmatched sample lines are excluded from figure 1 and from further discussion.

If a telephone number was found and the case was assigned to the **phone-first** condition, then the case was loaded into the computer-assisted telephone interviewing (CATI) system and dialed. Of the 4,689 sample lines assigned to phone-first, 1,508 (32.2 percent) finalized in CATI and never transitioned to mail. Most often, cases finalized in CATI when they completed the main interview or completed the screening interview and were determined to be ineligible. Some cases never changed modes because respondents refused in a hostile manner on the telephone (e.g., threatened legal action or used obscene language) or had other extenuating circumstances (e.g., a sudden death in the household). The remaining 3,181 (67.8 percent) phone-first cases were moved to the mail mode on a rolling basis. The

speed with which a case moved from CATI to mail depended on a complex set of calling rules defined within the CATI system.

If a telephone number was found and the case was assigned to the **mail-first** condition, then the case was mailed a paper SAQ (n = 4,800). Households in the mail-first condition that did not return a SAQ within 10 weeks (n = 3,222, or 67.1 percent) were moved to CATI. A mail return included cases in which a household member returned the SAQ or the mail was returned as undeliverable by the US Postal Service.

All operational procedures within mode, such as number and timing of mailings, calling rules, and screening procedures, were identical across the two groups. All dialed telephone numbers were asked to confirm their address to ensure that we were contacting the sampled address. Mailings for both groups followed the Dillman Tailored Design Method (Dillman, Smyth, and Christian 2009). An initial packet, including a $5 bill and two questionnaires, was mailed to each household. A reminder/thank-you postcard was mailed to all addresses after three weeks. A second packet (without an additional token payment) was mailed three weeks after the postcard to households that had not yet responded. All mail was sent using presorted standard postage through the US Postal Service.

Results

We evaluated the initial mode assignment and case flow on four survey metrics: response rates, cost, timeliness, and data quality. The comparisons were made throughout the life of a sample line, by both initial mode assignment and subsequent mode movement. Throughout the remainder of this paper, we refer to “phone-first” or “mail-first” to describe the case-flow design and “CATI” or “SAQ” to describe the mode of data collection. We were also interested in whether different population eligibility rates affect the differential performance of the case flow. Consequently, we present rates collapsed across and by three eligibility categories (high, medium, and low). All analyses were unweighted except for the final analysis on health statistics, for which we used base weights to adjust for probability of selection across communities. For all comparative analyses, we used standard t-tests or Bonferroni-adjusted t-tests to assess differences.

RESPONSE RATES

The initial goal of this experiment was to determine the approach that would maximize response. We calculated two rates for evaluating response. The screening response rate was calculated using American Association for Public Opinion Research (AAPOR) Response Rate 1 (2011). The interview completion rate is the number of interviews divided by the number of selected individuals within the household.[2] Table 1 contains response rates by mode of data collection within a case flow as well as combined response rates that are the final rates within a case flow. The combined rates are not the sum of the individual mode rates because the operational nature of the multimode design occasionally yields households

[2]Researchers often calculate a holistic response rate, combining the screening response rate and the interview completion rate. We did not calculate holistic response rates in this instance because multiple individuals were selected to complete the topical interview. This resulted in the screener response rate being calculated at the household level, while the interview completion rate was calculated at the individual level.

that complete the interview in more than one mode. Completion in more than one mode was more common among the sample lines assigned to the mail-first mode in which the SAQ may have been returned after we contacted the household by CATI. The combined response rates, therefore, required de-duplication to avoid double counting and inflating response rates.

The phone-first cases achieved significantly higher screener and interview response rates than the mail-first cases in CATI (25.8 percent versus 19.8 percent, $p < 0.0001$, and 55.2 percent versus 48.3 percent, $p = 0.003$, respectively), while the mail-first group obtained a significantly higher screener response rate via SAQ (32.6 percent versus 25.5 percent, $p < 0.0001$, respectively). The observed differences are not surprising given that the most willing respondents will participate regardless of data-collection mode. The initial data-collection mode captures willing respondents and so would be expected to fare better than if the same mode were used as the second mode.

The most important result was that the mail-first case flow had higher combined screener response rates and interview completion rates than the phone-first case flow. The screener response rate for mail-first was 3.9 percentage points higher ($p = 0.0003$), while the mail-first interview completion rate was 9.0 percentage points higher ($p < 0.0001$). Overall, households were more likely to eventually complete the interview if they initially received a SAQ by mail than if they were contacted initially via telephone in CATI.

One explanation is that the mail-first case flow may be less intrusive than phone-first. SAQ nonresponders who are later contacted via CATI may still be a “fresh” sample in that they do not remember receiving the SAQ packet or did not have to actively refuse to participate, simply ignoring the survey request. Thus, the decision to participate via CATI (the second mode) may be made independently of the decision on the first mode (SAQ). The phone-first case flow is likely a different response process than mail-first. If a respondent actively refused to participate via CATI and later received a mailed SAQ, he/she may remember the initial request and refusal and may be more likely to refuse the survey request for a second time. An alternative theory is that the SAQ acts as an advance letter and improves the efficiency of CATI. While we believe both of these hypotheses are reasonable, we do not have sufficient data to test them.

We believed that the survey eligibility rate, which is the proportion of screened households that have at least one individual who meets survey participation requirements, might influence the choice of optimal case flow. For comparative purposes, we classified the six communities into three eligibility groups. The “high”-eligibility group included three communities with an eligibility rate of 90 percent or higher. The “medium”-eligibility group contained two communities with eligibility rates ranging from 65 to 75 percent, and the “low”-eligibility group included one community with an eligibility rate of 25 percent. To put these rates in context, the high-eligibility group is similar to the expected eligibility rate achieved in a general-population survey; the medium-eligibility group is similar to a survey of registered voters; and the low-eligibility group is similar to a survey of households with children.

Table 2 reproduces the rates from table 1 across the three eligibility groups. The combined differences across groups were between 1.6 and 5.4 percentage points for the screener response rate and between 3.4 and 12.0 percentage points for the interview completion rate. Despite the overall trend of superior performance in the mail-first group, each eligibility group behaved somewhat differently. The high-eligibility group benefited most by the mail-first case flow, as it achieved a combined screener response rate 5.4 percentage points higher and a combined interview completion rate 8.8 percentage points higher than the phone-first case flow ($p = 0.0012$ and $p < 0.0001$, respectively). The medium-eligibility group also achieved higher rates for the mail-first case flow, with screener response and interview completion rate differences of 3.4 and 12.0 percentage points, respectively, but only the difference in the interview completion rate remained significant ($p = 0.095$ and $p < 0.0001$, respectively). For the low-eligibility group, there were even smaller differences—1.6 and 3.4 percentage point differences on the screener response rate and interview completion rate, respectively. However, neither difference was statistically significant ($p = 0.293$ and $p = 0.351$). Thus, in a mixed-eligibility survey of this type, the strong advantage of the mail-first case flow for high-eligibility populations may far outweigh the limited advantages in lower-eligibility populations.

COST

The higher response rates that resulted from the mail-first case flow are most valuable if they can be attained efficiently, as cost is often an important factor in initial mode assignment. To assess the efficiency of the case-flow paths, we calculated a variable cost ratio per completed interview. Variable costs per interview included interviewer labor and supervision; mailing costs such as printing and postage; receipting returned mail; and data entry for returned SAQs. The cost per completed interview was, then, the total variable cost divided by the number of interviews completed. A cost ratio of 1.0 suggests that interviews completed in either the mail-first or phone-first case flow would be of equal cost. A ratio below 1.0 implies that an interview from a mail-first case would be less expensive than a phone-first case. The reverse is true for a cost ratio above 1.0. Cost figures were only available in the aggregate by group, limiting the ability to conduct significance tests.

Overall, we found that the mail-first case flow provided a more cost-efficient model than the phone-first case flow (cost ratio = 0.88). Telephone interviewing costs were more expensive than mailing SAQs in aggregate, making the mail-first case flow more cost effective. Additionally, the mail-first case flow resulted in higher response rates, as noted above, further improving cost efficiencies.

We also analyzed costs by eligibility group to assess whether our initial assumption about a phone-first case flow in low-eligibility populations was supported with information about efficiency. In a general-population survey, every dollar spent on data collection has the potential to return a respondent, as nearly everyone is eligible to participate in the survey. As eligibility rates decline, however, more resources are spent screening out households. That is, more CATI interviewer labor hours are spent screening and more prepaid incentive money is lost to ineligible households. Therefore, screening costs may increase disproportionately by mode and thus affect cost efficiencies for the two case flows as

population eligibility declines. For completed interviews in the high-eligibility communities, the mail-first case flow cost 79 cents for every dollar spent for those in the phone-first case flow. In the low-eligibility category, each interview in the mail-first case flow cost 86 cents for every dollar spent to complete a phone-first case.

The medium-eligibility group departed from the trend with a combined cost ratio of 1.03. This suggests that interviews completed in the phone-first and mail-first case flows cost approximately the same for surveys in which population eligibility rates are between 65 and 75 percent of screened households. The higher cost ratio in this case was driven by the much higher costs of the CATI interviews conducted in the mail-first design. Given the small number of interviews conducted in CATI for both case flows in these communities (n = 193 and 105 for phone-first and mail-first, respectively), random variation in interviewing costs was likely responsible for this departure from the trend. Overall, we concluded that the combined data-collection costs of a mail-first case flow will be moderately lower than a phone-first case flow, regardless of the population eligibility rate.

TIMELINESS

The third metric for identifying an optimal case flow in this ABS multimode design is the length of the field period. Mail surveys usually require longer field periods, as it takes time to mail out an instrument and wait for a respondent to complete and return it, with returns ultimately occurring over several months. Although repeated attempts may be necessary, the timing of a telephone survey is usually at the discretion of the survey organization. Interviews conducted under the phone-first case flow took an average of 35.9 days to complete, whereas those collected under the mail-first design took an average of 48.1 days to complete ($p < 0.0001$). While this is not a nuanced idea, little work has been published on the amount of time needed to make a mail-first design more attractive than a phone-first case flow. The purpose of this analysis is to quantify the amount of time necessary in the field to achieve higher rates in the mail-first case flow.

We graphed the yield rate over time for the two case-flow methods, controlling for sample release date (figure 2). The yield rate is calculated as the total number of interviews divided by the fielded sample lines. The phone-first case flow achieved a higher sample yield rate faster, but it was surpassed by the mail-first case flow at two months into the data-collection period. The observed change in yield rates over time is driven by the higher yields overall in the mail-first case-flow design (as implied by the higher response rates observed in table 1). We conducted similar analyses by eligibility group (analysis not shown) and found very similar results.

The curve in figure 2 implies that if a data-collection field period exceeds two months, survey operations will benefit from a mail-first design. For shorter data-collection field periods, one should consider a phone-first case flow to capture the benefit from the faster initial yields. The two-month cutoff, is, however, influenced by the REACH U.S. mailing protocol in which presorted standard postage was used for all mailings. Under this protocol, mail is delivered within 7–10 business days (as opposed to 2–3 business days for first-class mail). Studies that use first-class or express-mailing procedures may see the inflection point

for sample yield between mail-first and phone-first designs approximately 1–2 weeks earlier in the field period than was observed in this experiment.

DATA QUALITY

Our final analyses focused on differences in data quality between the two case-flow models. Some research suggests that item nonresponse and mode differences between a self-administered questionnaire and one delivered by an interviewer may adversely impact the quality of data collected via SAQ (de Leeuw 2005; Dillman and Christian 2005). We used two indicators for data quality in our comparison. First, we examined item-level nonresponse rates across data-collection modes by case flow. The results presented in the previous sections suggest that a mail-first design may produce higher response rates at a lower cost. Researchers have previously found that interviews collected via mail result in higher item-level nonresponse (Dillman, Phelps, et al., 2009). More interviews will not be as useful if the additional cases have more missing data. Second, we evaluated the distributions of key statistics across data-collection modes by case flow. For both analyses, we evaluated demographic variables often used in weighting and nonresponse adjustment. We also analyzed different types of health measures, including a dichotomous measure of diabetes, a continuous measure of fruit servings per day, and a derived continuous variable, body mass index (BMI), created from the respondent's self-reported weight and height. The wordings of the questions used for both the demographic and health measure analyses may be found in appendix A. We also considered differences in the eligibility rate to identify differential nonresponse by case flow.

Item nonresponse—We considered a question to be missing if the respondent answered "don't know" or "refused" in CATI or left the question blank on the SAQ; "don't know" and "refused" options were not provided on the SAQ. Table 3 displays item-level nonresponse rates for a variety of demographic and health characteristics, with base-weighted percentages correcting for probabilities of selection across communities. The combined item-level nonresponse rates ranged from 0.8 to 8.0 percent. When all modes of data collection were combined, mail-first cases were less likely to provide responses for age ($p = 0.0103$), diabetes ($p = 0.0010$), and number of fruit servings consumed per day ($p < 0.0001$) than phone-first cases. Given the number of *t*-tests performed, a Bonferroni adjustment for multiple comparisons (not shown) was conducted. Even with the adjustment, the difference in fruit servings per day remained significant ($p < 0.0001$).

While we do not have a strong hypothesis for the results of age, the results of fruit servings were likely a function of the questionnaire format. Respondents were asked, "Not counting juice, how often do you eat fruit?" On the SAQ, respondents were provided a space to write a frequency and asked to mark a unit of measurement (i.e., per day, per week, per month, or per year). The higher item-level nonresponse is a function of more missing data for the unit of measurement. Respondents likely did not see or understand the additional requirement for this question and inadvertently skipped the unit of measurement.

Bias—Our second analysis evaluated differences in critical variable estimates. Although the mail-first design increased the overall response rate, a higher response rate does not

guarantee lower nonresponse bias (Groves 2006). We evaluated the two case flows by comparing the observed eligibility rate, demographic distributions, and health statistics. The unique survey geography of each REACH U.S. community makes it impossible to know the true population distribution of either demographic or health characteristics to which we could then compare estimates from the interviews completed through the two case flows. Most REACH U.S. communities do not conform to the same geographies used for BRFSS, the National Health Interview Survey (NHIS), or other surveys that would provide benchmarks for the demographic distributions or key health statistics. The surveys that do provide demographic benchmarks (e.g., the American Community Survey [ACS] or Current Population Survey [CPS]) do not report them in the small geographies by racial/ethnic subgroups and, even more restricting, for households for which a telephone number is available. Therefore, we do not have a benchmark standard for this analysis of potential bias by case flow. As an alternative, we compare the key estimates across the two case-flow mechanisms. While this is an imperfect alternative, it does provide an indication of whether the choice of starting mode will introduce additional bias into key measures. As with the item-level nonresponse analysis, estimates were weighted by the probability of selection to adjust for differences by REACH U.S. community.

There was little difference between the two case flows independent of how the interview was completed (table 4). The only exception was with regard to household income. Respondents living in households with an annual income less than $35,000 made up a larger proportion of total interviews in the mail-first design than the phone-first design (p = 0.0052). This difference, however, became nonsignificant after controlling for the number of comparisons using a Bonferroni adjustment. Research on mode effects suggests that respondents are more likely to report sensitive behaviors in a self-administered questionnaire (Fowler, Roman, and Di 1998). Based on the previous literature, we expected that diabetes prevalence and BMI estimates would be higher for the mail-first case flow and that reported daily fruit servings would be lower. The point estimates, however, were nearly identical by experimental group.

Discussion and Conclusion

In a case-flow experiment using data from REACH U.S., we found that the mail-first approach to a multimode design with an ABS sampling frame was superior to or on par with a phone-first design on all survey performance metrics. Screener response rates and interview completion rates were consistently and significantly higher for the mail-first design. These advantages persisted across communities with different survey eligibility rates. The largest response rate gains for the mail-first case flow were found in communities with nearly universal eligibility. Even in communities where only about a quarter of the population was eligible for the survey, a mail-first approach remained superior, albeit not statistically significant.

The other survey performance metrics we evaluated were not as unambiguously clear, but nevertheless reinforce the advantages of a mail-first case flow. In our analysis of variable costs, we found that, regardless of which mode the case completed, a mail-first case costs 86 cents for every dollar spent on a phone-first case. We again found that population eligibility

did not substantially alter the superiority of the mail-first design, although for the medium eligibility communities there was little difference between the two case flows. This was in part due to the small number of cases in each group and the resulting instability of the cost estimates. In our analysis of the length of the field period, we found that the mail-first case flow yielded more completed interviews, but it required a longer field period both within and between modes. Thus, mixed-eligibility surveys with a relatively long field period (i.e., two months or longer) will be well-served by an ABS multimode design that begins with a mail-first case-flow design.

The remaining issue was whether either case-flow design introduces differential bias by increasing population undercoverage, item nonresponse, or measurement error. Our analyses of data quality demonstrated very little difference between the two case flows. The only exceptions were higher item-level nonresponse rates for age, diabetes, and fruit intake in the mail-first design. Despite the relatively higher level of item-level nonresponse, nonresponse was acceptably low in both cases. Our analysis is not a definitive test of bias across designs because we did not have a benchmark to more rigorously assess bias. However, this analysis provides evidence that bias is not altered across the two designs.

There are several limitations to this study that may somewhat limit its generalizability. Most importantly, the experiment was limited to those sampled addresses that could be matched to a telephone number. In 2012, 59.8 percent of sampled addresses could not be matched to a phone number, making it impossible to field them in CATI. The magnitude of the differences isolated between the two case flows will diminish when evaluating the full sample. Despite the large number of sample lines that could not be fielded via telephone, the results of this experiment are still relevant to survey operations and design for several reasons. First, including the unmatched sample would likely lessen but not erase the importance of initial mode assignment. The unmatched sample could be fielded in the mail mode, but never in CATI. This will yield lower response rates than the matched sample that is fielded in multiple modes (table 1) and reduce the overall response rate. Even when accounting for the effect of the unmatched sample, the overall response rate will still be higher by fielding the matched sample via the mail-first case flow than by fielding the matched sample via phone-first. The magnitude of the difference will depend on the proportion of matched sample and the difference between the phone-first and mail-first rates. Using data from table 1 and the observed match rate of 40.2 percent in REACH U.S., we would expect to see an overall screener response rate 1.6 percentage points higher using a mail-first case flow for the matched cases (0.402*(0.487 – 0.448)) and an overall interview response rate 3.7 percentage points higher (0.402*(0.798 – 0.708)).

Second, the REACH U.S. address-to-telephone match rate is low due to the target geographies and populations and the technique we used to match addresses to telephone numbers. We encountered more cell-phone-only households, vacant addresses, and apartments than the national average. These types of addresses are less likely to be matched to a telephone number (Amaya, Skalland, and Wooten 2011). We also used a telephone matching algorithm that reduced the overall match rate in order to improve the accuracy of matches. Reverse address-to-telephone matching vendors provide a quality indicator that reflects how accurate the vendor believes the match is. Based on previous experience, we

opted to treat low-quality telephone number matches (often referred to as “inexact matches”) as unmatched samples and field them via the mail mode only. This approach lowered our overall match rate. Finally, the findings in this experiment likely apply to list frames where phone numbers and addresses may be available for all sample lines.

Additional research on both the starting mode and the justification and choreography of subsequent modes in an ABS multimode survey is necessary. Further experimentation that restricts mode movement as one of the experimental conditions may be appropriate. Comparisons across multimode and single-mode implementations for the same survey and the same population will allow for a more definitive answer to the value of a multimode design. Moreover, other subpopulation surveys should be analyzed for the same results. We focused on racial/ethnic groups, but different subpopulations may be differentially affected by mode order. The second avenue of research is to follow up more rigorously on our initial glimpse of the measurement consequences of a multimode design. Although the initial assignment to mode was random, mode movement occurred at later stages because of refusals or non-contacts. Population selection, which confounds mode comparisons, does not occur at the first stage but does occur as cases are shifted from the experimental assignment to the final mode for data collection. More sophisticated approaches to analyzing mode effects will be necessary to unravel the consequences of population selection from the impact of the mode of delivery.

Despite its limitations, this study has addressed a fundamental question about the most efficient case flow for an ABS multimode design. A mail-first approach, while requiring a longer field period, has clear and definitive advantages over a phone-first approach. It both yields higher response rates and is less expensive to execute. These advantages are consistent across all survey populations independent of their eligibility.

References

Amaya, Ashley; Skalland, Benjamin; Wooten, Karen. What’s in a Match? Survey Practice. 2010; 3:6.

American Association for Public Opinion Research. Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. 2011. Available at http://www.aapor.org/AAPORKentico/AAPOR_Main/media/MainSiteFiles/StandardDefinitions2011_1.pdf

Brick, J Michael; Andrews, WR.; Brick, Pat D.; King, Howard; Mathiowetz, Nancy A.; Stokes, Lynne. Methods for Improving Response Rates in Two-Phase Mail Surveys. Survey Practice. 2013; 5:3.

Brick, J Michael; Williams, Douglas; Montaquila, Jill M. Address-Based Sampling for Subpopulation Surveys. Public Opinion Quarterly. 2011; 75:409–28.

Couper, Mick. The Future of Modes of Data Collection. Public Opinion Quarterly. 2010; 75:889–908.

de Leeuw, Edith. To Mix or Not to Mix Data-Collection Modes in Surveys. Journal of Official Statistics. 2005; 21:233–55.

Dillman, Don A.; Christian, Leah M. Survey Mode as a Source of Instability in Responses across Surveys. Field Methods. 2005; 17:30–52.

Dillman, Don A.; Phelps, Glenn; Tortora, Robert; Swift, Karen; Kohrell, Julie; Berck, Jodi; Messer, Benjamin. Response Rate and Measurement Differences in Mixed-Mode Surveys Using Mail, Telephone, Interactive Voice Response (IVR), and the Internet. Public Opinion Quarterly. 2009; 38:1–18.

Dillman, Don A.; Smyth, Jolene D.; Christian, Leah M. Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method. 3. Hoboken, NJ: John Wiley; 2009.

Fahimi, Mansour. Enhancing the Computerized Delivery Sequence File for Survey Sampling Applications. Paper presented at the 65th Annual Meeting of the American Association for Public Opinion Research; Chicago, IL, USA. 2010.

Fowler, Floyd J.; Roman, Anthony M.; Di, Zhu X. Mode Effects in a Survey of Medicare Prostate Surgery Patients. Public Opinion Quarterly. 1998; 62:29–46.

Groves, Robert. Nonresponse Rates and Nonresponse Bias in Household Surveys. Public Opinion Quarterly. 2006; 70:646–75.

Groves, Robert. Three Eras of Survey Research. Public Opinion Quarterly. 2011; 75:861–71.

Iannacchione, Vincent. The Changing Role of Address-Based Sampling in Survey Research. Public Opinion Quarterly. 2011; 75:556–75.

Iannacchione, Vincent; Staab, Jennifer; Redden, David. Evaluating the Use of Residential Mailing Addresses in a Metropolitan Household Survey. Public Opinion Quarterly. 2003; 67:202–10.

Link, Michael W.; Battaglia, Michael P.; Frankel, Martin R.; Osborn, Larry; Mokdad, Ali H. A Comparison of Address-Based Sampling (ABS) versus Random-Digit Dialing (RDD) for General Population Surveys. Public Opinion Quarterly. 2008; 72:6–27.

Link, Michael W.; Daily, Gail; Shuttles, Charles D.; Christine Bourquin, H.; Tracie Yancey, L. Addressing the Cell-Phone-Only Problem: Cell Phone Sampling versus Address-Based Sampling; Proceedings of the Survey Research Methods Section of the Joint Statistical Meetings; 2008. p. 4229-4236.

Messer, Benjamin L.; Dillman, Don A. Surveying the General Public over the Internet Using Address-Based Sampling and Mail Contact Procedures. Public Opinion Quarterly. 2011; 75:429–57.

Millar, Morgan M.; Dillman, DA. Improving Response to Web and Mixed-Mode Surveys. Public Opinion Quarterly. 2011; 75:249–69.

Montaquila, Jill M.; Michael Brick, J.; Williams, Douglas; Kim, Kwang; Han, Daifeng. A Study of Two-Phase Mail Survey Data-Collection Methods. Journal of Survey Statistics and Methodology. 2013; 1:66–87.

Montaquila, Jill M.; Hsu, Valerie; Michael Brick, J. Using a ‘Match Rate’ Model to Predict Areas Where USPS-Based Address Lists May Be Used in Place of Traditional Listing. Public Opinion Quarterly. 2011; 75:317–35.

Montaquila, Jill M.; Valerie Hsu, J Michael; Brick, Ned English; O’Muircheartaigh, Colm. A Comparative Evaluation of Traditional Listing vs. Address-Based Sampling Frames: Matching with Field Investigation of Discrepancies. Proceedings of the Survey Research Methods Section of the Joint Statistical Meetings. 2009:4855–4862.

Murphy, Whitney; Harter, Rachael; Xia, Kanru. Design and Operational Changes for the REACH U.S. Risk Factor Survey. Proceedings of the Survey Research Methods Section of the Joint Statistical Meetings. 2010:332–42.

O’Muircheartaigh, Colm; Eckman, Stephanie; Weiss, Charlene. Traditional and Enhanced Field Listing for Probability Sampling. Proceedings of the Survey Research Methods Section of the Joint Statistical Meetings. 2003:2563–2567.

Williams, Douglas; Montaquila, Jill M.; Michael Brick, J.; Hagedorn, Mary C. Screening for Specific Population Groups in Mail Surveys. Paper presented at the 65th Annual Meeting of the American Association for Public Opinion Research; Chicago, IL, USA. 2010.

Appendix A. REACH U.S. Wording of All Questions Used in Data Analysis

EDUCATION. What is the highest grade or year of school you completed?

(RESPONSE OPTIONS IN SAQ AND DISPLAYED TO CATI INTERVIEWERS.)

Never attended school or only attended kindergarten; Grades 1 through 8 (Elementary); Grades 9 through 11 (Some high school); Grade 12 or GED (High school graduate); College 1 year to 3 years (Some college or technical school);

College 4 years or more (College graduate); Don’t know (CATI only); Refused (CATI only)

AGE. What is your age?

(OPEN-ENDED RESPONSE)

INCOME. (For SAQ) Is your annual household income from all sources…?

Less than $10,000; $10,000 to less than $15,000; $15,000 to less than $20,000; $20,000 to less than $25,000; $25,000 to less than $35,000; $35,000 to less than $50,000; $50,000 to less than $75,000; $75,000 or more (For CATI) Is your annual household income from all sources…

Less than $25,000?

Less than $20,000?

Less than $15,000?

Less than $10,000?

Less than $35,000?

Less than $50,000?

Less than $75,000?

Yes; No; Don’t know; Refused

(SKIP LOGIC AS APPROPRIATE.)

DIABETES1. Have you ever been told by a doctor that you have diabetes?

(RESPONSE OPTIONS IN SAQ AND DISPLAYED TO CATI INTERVIEWERS.)

Yes; No; No, pre-diabetes or borderline diabetes; Don’t know (CATI only); Refused (CATI only)

DIABETES2. (If DIABETES1 = YES and GENDER = FEMALE) Was this only when you were pregnant?

(RESPONSE OPTIONS IN SAQ AND DISPLAYED TO CATI INTERVIEWERS.)

Yes; No; Don’t know (CATI only); Refused (CATI only)

BMI1. About how much do you weigh without shoes?

(RESPONSE ALLOWED IN LBS OR KILOS AND CONVERTED TO LBS FOR ANALYSIS.)

BMI2. About how tall are you without shoes?

(RESPONSE ALLOWED IN FT/IN OR CM AND CONVERTED TO INCHES FOR ANALYSIS.)

FRUIT. Not counting juice, how often do you eat fruit?

(RESPONSE ALLOWED PER DAY, PER WEEK, PER MONTH, OR PER YEAR AND CONVERTED TO PER DAY FOR ANALYSIS.)

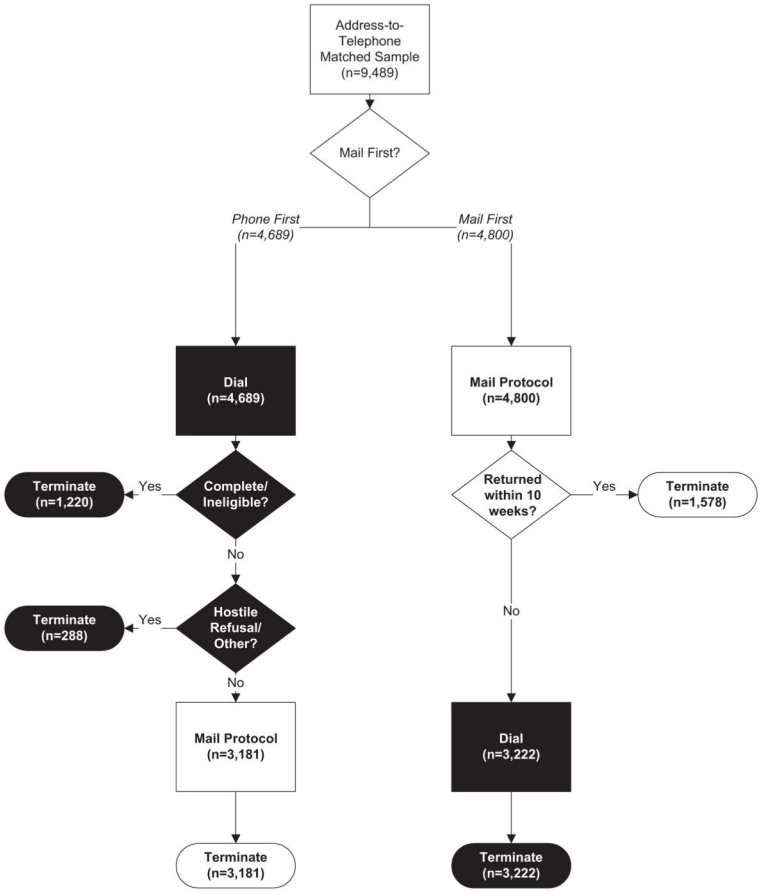

Figure 1. REACH U.S. Case Flow by Experimental Condition

Sample sizes include only cases with an exact telephone match: 59.8 percent of sample lines in the six communities did not have a telephone match and are excluded from this analysis and the case counts listed here.

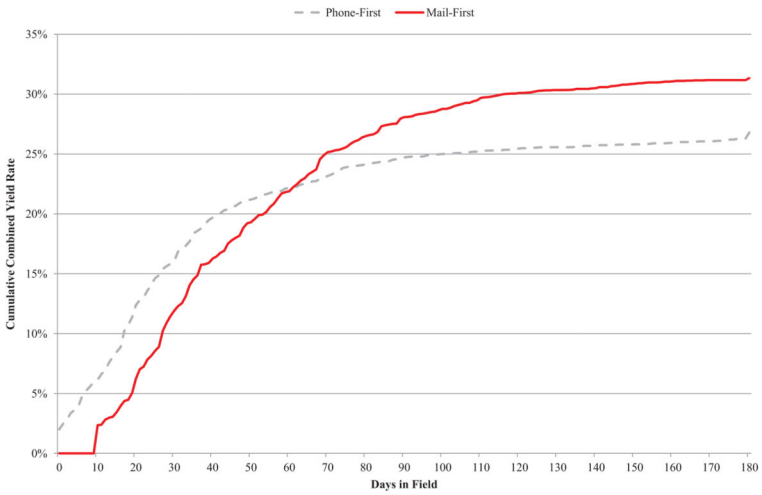

Figure 2.
Cumulative Combined Yield Rate by Days in Field and Case Flow.

Table 1

Response Rates by Data-Collection Mode and Case Flow

	Case flow		
Data-collection mode	**Phone-first**	**Mail-first**	***p*-value**
CATI			
N	4,689	3,222	
Completed topical interviews	657	370	
Screener response rate	25.8	19.8	< 0.0001
Interview completion rate	55.2	48.3	0.0030
SAQ			
N	3,181	4,800	
Completed topical interviews	609	1,183	
Screener response rate	25.5	32.6	< 0.0001
Interview completion rate	100.0	99.8	N/A
Combined			
N	4,689	4,800	
Completed topical interviews	1,266	1,537	
Screener response rate	44.8	48.7	0.0003
Interview completion rate	70.8	79.8	< 0.0001

Note.—Regarding the SAQ interview completion rate, the REACH U.S. mailing included the screening and topical interviews in the same packet. Nearly all cases that completed and returned the screening interview also completed the main interview, hence the high interview completion rate. REACH U.S. coded a SAQ as complete if the respondent had answered at least the first question from every section of the survey and had provided enough information to determine respondent race/ethnicity and age category. *P*-values could not be calculated for the SAQ interview completion rate since there was no variation among the phone-first cases.

Table 2

Response Rates by Data-Collection Mode, Case Flow, and Eligibility Group

	Eligibility group by case flow								
	High			Medium			Low		
Data-collection mode	**Phone-first**	**Mail-first**	***p*-value**	**Phone-first**	**Mail-first**	***p*-value**	**Phone-first**	**Mail-first**	***p*-value**
CATI									
N	1,899	1,353		1,193	891		1,597	978	
Completed topical interviews	385	230		193	105		79	35	
Screener response rate	23.9	22.1	0.2365	22.9	16.2	0.0002	30.1	19.9	< 0.0001
Interview completion rate	53.3	46.3	0.0167	55.5	53.6	0.7176	65.8	47.3	0.0107
SAQ									
N	1,306	1,835		858	1,258		1,017	1,707	
Completed topical interviews	337	664		180	347		92	172	
Screener response rate	21.5	27.3	0.0002	21.3	29.4	< 0.0001	34.2	40.7	0.0008
Interview completion rate	100.0	99.7	N/A	100.0	99.7	N/A	100.0	100.0	N/A
Combined									
N	1,899	1,835		1,193	1,258		1,597	1,707	
Completed topical interviews	722	885		373	447		171	205	
Screener response rate	39.7	45.1	0.0012	40.4	43.8	0.0950	54.7	56.3	0.2930
Interview completion rate	68.6	77.4	< 0.0001	71.1	83.1	< 0.0001	81.0	84.4	0.3508

Note.—Regarding the SAQ interview completion rate, the REACH U.S. mailing included the screening and topical interview in the same packet. Nearly all cases that completed and returned the screening interview also completed the topical interview, hence the high interview completion rate. REACH U.S. coded a SAQ as complete if the respondent had answered at least the first question from every section of the survey and had provided enough information to determine respondent race/ethnicity and age category. *P*-values could not be calculated for the SAQ interview completion rate since there was no variation among the phone-first cases.

Table 3

Item-Level Nonresponse Rates by Data-Collection Mode and Case Flow

	Case flow		
Data-collection mode	**Phone-first**	**Mail-first**	***p*-value**
CATI			
N	657	370	
Education	0.4	0.6	0.3723
Age	0.8	0.3	0.0693
Income	7.8	7.3	0.5663
Diabetes	2.5	3.3	0.4753
BMI	2.5	6.0	0.0112
Fruit servings per day	0.2	0.3	0.7071
SAQ			
N	609	1,183	
Education	1.3	1.0	0.2011
Age	0.8	2.4	< 0.0001
Income	7.3	8.2	0.1907
Diabetes	6.9	8.8	0.1534
BMI	3.5	3.2	0.7861
Fruit servings per day	5.3	7.3	0.0893
Combined			
N	1,266	1,537	
Education	0.9	0.9	0.9054
Age	0.8	1.9	0.0103
Income	7.6	8.0	0.6689
Diabetes	4.6	7.5	0.0010
BMI	2.9	3.9	0.1548
Fruit servings per day	2.6	5.6	< 0.0001

Table 4

Key Variable Distributions by Data-Collection Mode and Case Flow

	Case flow				
	Phone-first		Mail-first		
Data-collection mode	**Percent**	**CI**	**Percent**	**CI**	***p*-value**
CATI					
N	657		370		
Eligibility rate	43.0	40.1–45.8	57.1	53.2–61.0	< 0.0001
High school education or less	54.2	50.3–58.0	57.9	52.8–63.0	0.2243
Age (mean)	56.4	55.0–57.7	55.6	53.7–57.4	0.4808
Household income below $35k	59.5	55.6–63.5	70.0	65.1–74.9	0.0016
Diabetes	23.5	20.2–26.8	23.8	19.3–28.2	0.9191
BMI (mean)	29.7	29.1–30.2	30.5	29.7–31.3	0.0877
Fruit servings per day (mean)	3.7	3.6–3.9	3.7	3.5–4.0	0.9654
SAQ					
N	609		1,183		
Eligibility rate	42.7	39.3–46.1	37.4	35.0–39.8	0.0123
High school education or less	46.9	42.9–50.9	46.7	43.8–49.6	0.9129
Age (mean)	52.9	51.6–54.2	53.6	52.7–54.6	0.3661
Household income below $35k	56.4	52.3–60.5	61.5	58.6–64.3	0.0480
Diabetes	21.1	17.7–24.5	22.2	19.7–24.6	0.6273
BMI (mean)	30.5	30.0–31.0	29.9	29.5–30.3	0.0664
Fruit servings per day (mean)	3.8	3.5–4.1	3.8	3.6–4.0	0.7876
Combined					
N	1,266		1,537		
Eligibility rate	42.8	40.7–45.0	42.3	40.2–44.4	0.7277
High school education or less	50.3	47.5–53.0	49.7	47.2–52.2	0.6887
Age (mean)	54.5	53.6–55.5	54.0	53.2–54.9	0.4247
Household income below $35k	57.8	55.0–60.6	63.3	60.8–65.8	0.0052
Diabetes	22.2	19.8–24.5	22.4	20.2–24.6	0.8754
BMI (mean)	30.1	29.7–30.3	30.0	29.6–30.3	0.6085

	Case flow				
	Phone-first		Mail-first		
Data-collection mode	Percent	CI	Percent	CI	*p*-value
Fruit servings per day (mean)	3.8	3.6–3.9	3.8	3.6–3.9	0.9135