

## Survey Practice June 2009 Issue

Monday, June 29, 2009, 7:17:09 AM | Editor

Welcome to Survey Practice in a new format. In the new format:

- \* Survey Practice will be published bimonthly;
- \* Has 6-7 articles in each bimonthly issue, instead of 3-4 in each monthly issue; and
- \* Each issue has an introduction and table of contents instead of first paragraph of each article.

The next issue will be published in August and will be a special topics issue on the uses of non-probability samples in survey research.

Our “Ask the Experts” article this month is the second by Aaron Maitland on the labeling of attitude scales. His article in the previous issue of Survey Practice on [labeling scale points](#) was the most viewed by SP readers. This month, his article focuses on [the number of responses to include in attitude scales](#). Please let the editors know if you would like more articles like these.

Prior to the last issue of Survey Practice, the article most viewed was on the [universal design for surveys](#). In this issue, Larry Malakhoff provides more detail on some requirements needed to be “508” compliant. That is, he describes some [formatting and other design issues](#) that web survey developers need to use to meet the federal guidelines for accessibility.

The hottest topic at this year’s AAPOR annual meeting was [address-based sampling](#). Many think that it will replace telephone samples and that it has the flexibility to be used for paper and Web surveys, too. We asked the presenters at the conference to prepare a short summary of their presentations. Some presenters provided longer papers or PowerPoint slides that are available at the end of their summary. ABS is an interesting and important method and more research is needed. The SP editors are considering a special issue on ABS sometime in the next year.

The [raking article](#), written by Mike Battaglia and his colleagues, is a little longer than most SP articles. But, the topic is interesting and requires more than the standard number of words to convey its message. Survey Practice often has articles that might seem relevant to a limited group of readers but we also hope that others will read the articles to learn more about a topic that might be a stretch for them. This article is a good non-technical description of raking.

The article by Katherine McGonagle and her colleagues at the University of Michigan describes an [experiment using a variety of methods to maintain the panel members](#) of the Panel Study of Income Dynamics. They found some difference among methods that would be useful for others who need to maintain panels.

Heather Stuckey and her colleagues present a good analysis of [the cost-effectiveness of pre-survey notifications and return receipts](#). Many of us often wonder if extra efforts to improve efficiency and response rates are justified by the extra costs. This study helps to understand the tradeoffs and shows that their use of a pre-survey postcard mailing saved money.

Survey Practice accepts articles in multiple formats but most articles sent to us are in traditional magazine or journal format. The article by Glenn Israel is an example of short article that describes useful research. He shows [mode differences in response between a postal and web survey](#). The article can be used as an example for others who would like to submit short, descriptive articles.

And, as always, we welcome your comments and suggestions. Let us know what kinds of articles you would like to see in SP. Do you prefer the monthly or bimonthly format?

### Articles in the Survey Practice June 2009

[How Many Scale Points Should I Include for Attitudinal Questions?](#)

[Can Survey Respondents with Visual Deficits Complete My Web Survey?](#)

[Summaries of Address-Based Sampling Presentations at the AAPOR Annual Meeting](#)

[Considerations in Raking Survey Data](#)

[An Experimental Test of a Strategy to Maintain Contact with Families between Waves of a Panel Study](#)

[Analyzing the Cost-Effectiveness of Using Return Receipt and Address Corrections in Mail Surveys](#)

[Obtaining Responses by Mail or Web: Response Rates and Data Consequences](#)

The Editors

John Kennedy            Diane O'Rourke  
David Moore            Andy Peytchev

[survprac@indiana.edu](mailto:survprac@indiana.edu)

Posted in Monthly Summary Comments: 0

[Comments \(0\)](#)

## [How Many Scale Points Should I Include for Attitudinal Questions?](#)

Monday, June 29, 2009, 7:15:49 AM | Editor

Aaron Maitland  
National Center for Health Statistics\*

Response scales are frequently used to measure attitudes in survey research. In this short article, I will discuss some theoretical considerations in the measurement of attitudes that influence the number of scale points. Next, I will discuss empirical results regarding the quality of scales with different numbers of scale points. Last, I will provide some question development strategies.

There are various theoretical considerations for determining the number of scale points. Attitudes are abstract constructs that are not directly observable and exist only in the respondent's mind. Response scales allow respondents to express both the direction and intensity of their attitudes. Some attitudes are viewed as bipolar concepts where two opposing sides of a concept are measured, whereas other attitudes are viewed as unipolar concepts where only the level of an attitude or just one side of a concept is measured. Researchers must clearly define the attitude object towards which a respondent can express an attitude. The respondent can then represent his or her stance for or against an attitude object by selecting the appropriate option on the response scale. Researchers usually conceptualize attitudes as existing along an attitude continuum. Hence, response scales that allow respondents to express different shades of an attitude rather than being simply for or against an attitude object will allow for better measurement of that continuum.

However, the difficulty of the response task must also be considered when designing response scales. Although longer scales might seemingly measure the attitude continuum in more detail, the response task might become too demanding, with too many scale points. This might force respondents to make more finely graded distinctions between scale points than might be possible. Respondents then might decide that the question is too demanding and *satisfice* (Krosnick 1991) by choosing the first plausible option that they encounter rather than carefully considering all options along the scale (Krosnick and Fabrigar 1997). Additionally respondents might resort to rounding their answers.

It is also important to consider the mode in which a question is going to be administered. The key distinction is between modes that rely solely on oral communication such as telephone versus modes that make use of visual communication such as the Web, mail, or face to face surveys with show cards. Although it depends on a number of factors such as how many scale points are labeled, one might generally conclude that longer response scales are easier to administer in a mode that uses visual communication since the respondent does not have to store all of the options in memory. The advantage of visual communication is probably minimal if only the endpoints of a lengthy response scale are labeled. Furthermore, unfolding or two step procedures can be used in telephone surveys to offer more response options without forcing the respondents to store the full range of options in short-term (working) memory. Research has found very few differences between the answers to questions using this unfolding technique over the telephone and those administered with a show card in face to face surveys (Groves 1988).

Finally, one must consider the interpretability of a middle position and whether it is meaningful for a specific concept. One interpretation is that respondents use this option when the middle category accurately describes their position (i.e., neither for nor against). Others suggest that a middle position is often interpreted as a “no opinion” option or an invitation to take an easy out for respondents who actually do have opinions, but are either unwilling or unable to express them due to the cognitive burden of the survey question (Krosnick 1991).

There are data quality standards that can be used to provide some insight into the optimal number scale points. Reliability and validity are two data quality standards most often employed using a quantitative framework. Reliability refers to how consistent answers are over replications. Reliability is measured over replications of the same question at different points in time or over multiple questions measuring the same attitude on a single occasion. Validity in the context of attitude measurement refers to how closely a survey question measures the construct of interest. Validity is difficult to measure, but is often operationalized quantitatively by assessing the extent to which a question converges with other questions measuring similar constructs and diverges from other questions measuring different constructs (Saris and Gallhofer 2007). Qualitative research methods are also useful for assessing the quality of survey questions. For example, in depth *cognitive* interviews can provide a detailed understanding of how survey respondents use the response categories and allow the researcher to assess whether this matches the question designer’s intent.

Several empirical studies have examined the effect of the number of scale points on the reliability of questions with response scales. The literature is mixed, probably indicating that the number of scale points depends on the specific objectives of a research project. Nonetheless, I will highlight some of the important conclusions that have been drawn. A review by Krosnick and Fabrigar (1997) did not find a monotonic increase in reliability as the number of scale points increased. Instead, a curvilinear pattern emerged in their review such that scales between 5-7 points were more reliable than scales with fewer points or more points. This was true for both bipolar and unipolar scales. Another study analyzed the longitudinal reliability of more than 300 survey questions that were repeated at more than two points in time (Alwin 2007). Once again, there was no monotonic increase in reliability as the number of scale points increased. Overall, two point scales were the most reliable followed by four, five, and nine point scales. Reliability was lower for six and seven point scales. The high reliability for two point scales could be due to the fact that two point scales only measure direction, whereas larger scales measure both direction and intensity (Alwin and Krosnick 1991). Interestingly the results were clearer for unipolar than bipolar scales. Four and five point unipolar scales demonstrated superior reliability compared to unipolar scales of other lengths; however, there were smaller differences in reliability between bipolar scales of different lengths.

There is some evidence that a middle position leads to lower reliability for shorter scales. Alwin (2007) found that three category scales are less reliable than two or four category scales. However, there was no clear evidence that five point scales were less reliable than four or six point scales. This suggests that the damaging effect of a middle position on reliability weakens as the number of categories increase.

The issue of validity has been addressed less frequently in the literature. Krosnick and Fabrigar (1997) report evidence that supports their view that 5-7 scale points are optimal. They found that questions using scales in this range tended to correlate the strongest with questions measuring conceptually related variables. Another interesting finding is that context effects – the effects of previous questions on a target question – tend to weaken as the number of scale points increases up to 7 points, after which there is very little change. They also report that even though the proportion of scale points used stays fairly constant up to 19 points, scales longer than 7 points do not seem to convey any additional information to researchers.

Although much of the evidence does seem to converge around the conclusion that 5-7 points might be optimal, others argue that more scale points is better. Based on results from multi-trait, multi-method experiments, Saris and Gallhofer (2007) conclude that up to 11 categories may be optimal. They claim that other authors are mistakenly interpreting variation from longer scales as measurement error. In short, they argue that different people's attitudes are calibrated differently so that similar opinions might be expressed with different values. For example, some respondents might have a tendency to express themselves with extreme words, whereas others express themselves more moderately. This issue becomes more pronounced with longer scales. To prevent variation due to these individual response differences the authors argue for the use of fixed reference points (e.g., completely disagree, completely agree) at the end points of a scale to help reduce this type of variation.

Despite the principles that have been discussed in this article, many issues are unresolved and the choice of response scales should be driven by research objectives. Pretesting enables researchers to match question design with these objectives. In-depth, qualitative, or *cognitive* interviews making use of think-alouds or probing techniques can help a researcher understand if a response scale resembles how respondents tend to think about and answer survey questions. In other words, these in-depth interviews should lead to better validity by more closely matching the response scales with the respondents own representations of an attitude. This technique also provides an understanding of the burden that a scale places on respondents. Given enough time and budget, field experiments that include different forms with scales of different lengths are another useful technique. Additionally, it is important to include repeated measurements of the response scales within a survey to assess their reliability.

\*The findings and conclusions in this report are those of the author and do not necessarily represent the views of the Centers for Disease Control and Prevention.

## References

Alwin, D.F. *Margins of Error: A Study of Reliability in Survey Measurement*. New York, NY: John Wiley and Sons, Inc. 2007.

Alwin, D.F., and Krosnick, J.A. "The Reliability of Survey Attitude Measurement: The Influence of Question and Respondent Attributes." *Sociological Methods and Research* 20 (1991): 139-181.

Groves, R.M. *Telephone Survey Methodology*. New York, NY: John Wiley and Sons, Inc. 1988.

Krosnick, J.A. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology* 5 (1991): 201-219.

Krosnick, J.A., and Fabrigar, L.R. "Designing Rating Scales for Effective Measurement in Surveys." *Survey Measurement and Process Quality*. Ed. Lyberg, L., Biemer, P., Collins, M., de Leeuw, E., Dippo, C., Schwarz, N., Trewin, D. New York, NY: John Wiley and Sons, Inc. 1997. 141-164.

Posted in Ask the Experts, Methods [Comments: 0](#)

[Comments \(0\)](#)

## [Can Survey Respondents with Visual Deficits Complete My Web Survey?](#)

Monday, June 29, 2009, 7:14:17 AM | Editor

Lawrence A. Malakhoff, U.S. Census Bureau

It is another ordinary day at the survey research center, and a Web survey designer is drinking coffee while reading over a request for a proposal to conduct a Web survey. The request states that because the sponsor is a U.S. Government agency, the Web survey must be accessible and must conform to Section 508 of the Americans with Disabilities Act.

Conformance with Section 508 is an unfamiliar requirement, so the Web survey designer investigates by typing "Section 508" in the Google search window. The first ten of 13.8 million results appear. The results page offers some links which reference the standard and others that provide checklists. The designer needs to understand this requirement before the survey research center can bid on the task, let alone conduct the survey if their bid is accepted.

### **Defining Accessibility**

The core principle in creating accessible and usable software for persons with disabilities is equivalent access. Information must be accessible to all users. This does not mean *equal* access. For example, when some software is installed, sighted users will see a progress bar moving to the right with a percentage value being updated continuously until it reaches 100%. A person with a visual deficit using a screen-reader would hear an announcement every few seconds about completion status – "50 percent complete ... 80 percent complete." Information about installation progress provides constant updates to persons who can see, but periodic updates to screen-reader users because progress changes faster than it can be announced. Similarly, images can be made accessible by creation of a short text description known as alternate, or ALT, text. Therefore, a screen-reader announcing progress toward installation completion and reading image ALT text satisfies the requirement for equivalent access.

## The Regulation Versus the Intention of Section 508

Accessibility is not a new requirement and has been a U.S. Government regulation since June 2001. Automated accessibility testing tools are available. Free tools like Cynthia Says [1] test just one page at a time. Tools such as AccVerify[2] and InFocus [3] can check entire Web sites and must be purchased.

\*\*\*

Since future work for the U.S. Government will require accessible software, the Web survey designer decides to invest in an automated testing tool. The next question is how to interpret the results. The output of the tool lists accessibility violations and provides the line number in the code where the violation occurred. As the Web survey designer reads about using the automated tools, the literature indicates that such tools will interpret each and every accessibility guideline literally, without consideration to what else is on the page [4]. The designer gets the feeling that these tools do not exhibit a lot of common sense.

\*\*\*

A Web application may be coded perfectly and conform to a strict coding structure, but if the content is poorly structured, the software will be difficult or impossible to use for persons with visual deficits. Content structure is the most important part of accessibility. It is possible to create an application that will pass the automated accessibility tests, but be unusable and therefore inaccessible.

Windows PC users who are blind or have low vision often use screen-reader software, either Window-Eyes [5] from GW Micro or Job Access With Speech (JAWS) from Freedom Scientific [6]. Screen-reading software permits users to hear text content of Web pages and Windows applications. Web survey designers need to realize screen-reader users will require more time than mouse users do to access form fields, questions, and navigation buttons, and that screen reader users can only manage large text blocks if sections are marked with text headers. For screen-reader users, making sense of unstructured text is like trying to drink water from a fire hydrant with a straw — there is just too much information.

## The Top 5 Accessibility Issues

During a five year period (2004-2009), the author reported on [accessibility issues](#) for software applications submitted to the Usability Lab in the Statistical Research Division at the U. S. Census Bureau. Navigation problems accounted for fourteen percent of the accessibility issues; these navigation issues were divided between incorrectly programmed tab order and problems with skip link navigation, which is used by screen-readers users to bypass groups of repetitive links to access content quickly. Fifty-four percent of the accessibility issues were related to missing or incorrect text. Web survey designers must create meaningful ALT text for any graphic; ensure plain text is accessible and not an image of text; and use descriptive labels for data-input fields. If ALT text, plain text, and labels are accessible on a Web survey form, then it will be possible for a screen-reader user to complete a Web survey.

## Types of Web Surveys

Web surveys may be implemented in one of two ways: screen-based or scrolling/paging. A screen based form has one question per screen. The user navigates by using previous/next buttons. The second method is a scrolling or paging implementation, in which all questions are on one page, and the user navigates by scrolling up or down the form. Either implementation method can be made accessible, but a screen-based form has the advantage of delivering text in more manageable blocks, and providing the Web survey designer the benefit of using a question response to customize the phrasing of any following questions. For example, on a screen-based form, if a respondent answers all children living there are in boarding school, later questions about them using the public bus to get to school could be skipped.

## Usage of Color

Paragraph I, 1194.21 of Section 508 [7] states: “Color coding shall not be used as the only means of conveying information, indicating an action, prompting a response, or distinguishing a visual element.” Therefore, color alone cannot be used to specify an action or convey any information required by a user to accomplish a task. For example, if a screen has a light green button labeled with “GO” in dark green text, instructions must not rely on color. Since users with a color deficit see in shades of gray, a statement like “Press the green button” would not help the user accomplish the task. Instead, instructing the user to press the GO button makes it possible for persons with a color deficit to successfully complete the task.

## Visual Focus and Reading Order

Screen-reader software will vocalize text beginning on the upper left part of a screen and proceeding down to the lower right, following the visual reading order. Reading order may be modified by changing tab number values so text can be read down two adjacent columns (column 1, then column 2) instead of left to right and top to bottom. Reading order may need to be modified to match the visual order. As a user tabs through the interface, the application must show visual focus, per Paragraph C, 1194.21 of Section 508 [7]. Visual focus is accomplished by showing a box around a screen element that moves to the next element when the tab key is pressed.

Screen-reader users must navigate backward to hear content on the top of a screen and forward to access content towards the bottom of a screen. For this reason, persons using screen-reader software are said to have **linear access**. Mouse users have **random access** because they can click on any screen element immediately without needing to step through other areas on the screen.

## Memory Burden

Access to application screens presents a greater memory burden to screen-reader users than to persons with normal vision because information preceding or following the current cursor location is not immediately available. More text present on the screen requires more time to navigate to the relevant text, such as a question, and raises the risk a screen-reader user may

forget the question when accessing the data-input field. Questions should be structured to accommodate short-term memory issues. Unless the material is transferred to working memory, the first items mentioned tend to be the first items forgotten [8]. Sighted users do not have this problem because they can scan back to instructions easily.

Placement of Instructions On the left side of [Figure 1](#) the question is listed first, followed by instructions, then a response field. The right side of Figure 1 contains a topic phrase to set up context for the directions, which follow next, after which the question and a response options area are provided. Lengthy instructions can interfere with recall of the original question. Screen-reader users will likely have better success in remembering the question if there are no instructions between the response field and the question.

Stem-and-Leaf Structure Screen-reader users may find questions using a stem-and-leaf structure challenging to comprehend. A question stem contains the first part of a question, followed by two or more conditions (the second part of the question) called leaves. Listening to the first leaf does not present a problem because it follows the stem. The issue arises if a screen reader user is positioned on the second and later leaves and cannot remember the stem text; then backward navigation is needed to hear the stem text again. Stem-and-leaf structures are technically accessible, but have poor usability.

Inferences and Memory Burden Inferences can be used to streamline a Web survey questionnaire. The amount of text and user burden could be reduced if question wording is inferred from earlier responses. [Figure 2](#) shows that if a respondent chose condition 2, they should not hear instructions for conditions 1 and 3 in the following question. [Figure 3](#) demonstrates that inferences can also be used to make Web surveys less impersonal. If gender is asked, this information could be used to word later questions with personal pronouns (he, she, his, hers). Mispronounced words, such as first names, may cause screen-reader users to back-track to hear the name again, so personal pronouns are preferable to first names because they will not be mispronounced by the screen reader.

## Navigation Instructions

Web survey designers must consider the differences in perception between screen-reader users and persons with normal vision. Specifically, in the case of screen layout, an instruction contained in a column that tells users to choose a link to the left or right poses a barrier to screen-reader users. While the link appears visually to be a short distance away from the instruction on the screen, the screen-reader user must first guess which direction to navigate, then press the tab key a number of times to access the link. This situation can be avoided if a link always immediately follows the direction to select it in the tab order. If a link is accessible, a direction to “select the link below” usually will not pose problems for screen-reader users because they can infer they must navigate forward to access the link.

## Testing Methodology

Automated accessibility testing on a Web survey should be performed on a Windows PC with speakers. Testing tools and screen readers are available for other operating systems, but

Windows users dominate the market. Unless the Web survey is a scrolling form, it is likely that only one screen can be tested at a time. The automated accessibility testing software does a good job at identifying missing labels for graphics and buttons but does nothing to check for a logical tabbing order or whether the ALT text for a graphic makes sense.

An example showing correct reading order is provided in [Figure 4](#). Tabbing order and readability are best checked with a screen-reader. The tester should obtain a free demo copy of either the Window-Eyes [\[5\]](#) or JAWS [\[6\]](#) screen-reader. Launch the Web survey, start the screen-reader, and begin tabbing through the survey. Check the text between headers and buttons with the up and down arrow keys. The demo software runs for 30 minutes (Window-Eyes) or 40 minutes (JAWS) before the tester must reboot the PC.

### **Recommendations for Accessible and Usable Web Surveys**

Make it usable – Usability testing needs to be done with test participants performing typical tasks, and any issues that are observed during testing need to be corrected. The Web survey should be easy to use and allow users to provide their information accurately and efficiently, with high satisfaction. Minimize memory burden, and avoid stem-and-leaf questionnaire structures. Remember: If it's not usable, it's really not accessible. The Web site [usability.gov](http://usability.gov) offers guidance on creating usable software [\[9\]](#).

Make it accessible – An automated accessibility testing tool and screen-reader software should be used to evaluate the Web survey for conformance to Section 508 and to assess equivalent access. Color should not be used alone to convey information. Software applications must adhere to Section 508 Web guidelines to be accessible. [\[7\]](#)

Ensure correct reading order – The screen-reader user must experience the same visual sequence of questions, answer choices, skip patterns and instructions as that experienced by the sighted user. Does the text read in the same order when accessed by tabbing/arrow keys as a sighted person would read it? Does the application show focus (a box around a screen element) when the tab key is pressed? Even though a questionnaire screen may appear to be laid out properly, the text still may not be read aloud in the correct order by screen-reader software. Screen elements are automatically assigned a tab number when they are placed onto an application page, which may not coincide with the visual reading order. Web designers can revise the tab order with their development software to ensure the correct reading order. Jim Thatcher [\[10\]](#) and WebAIM [\[11\]](#) both offer guidance on Web form layout.

Create a single accessible version – It is more efficient to create one accessible application rather than two versions, such as “text only” and “graphical (point and click)”. Frequently, “text only” versions of software are not updated when the graphical version is modified. A text alternative should only be considered as a last resort and should be updated whenever the graphical version changes.

Integrate accessibility into the design process – Accessibility should be part of the design process from the beginning. Adding in accessibility in the final stage of development can be costly and

is not always possible. Design modifications for accessibility should occur at the beginning of the process before programming begins when changes are easier to make.

## Summary

There is much for Web survey designers to keep in mind when designing surveys to be Section 508 compliant. It is important to remember that accessible text does not guarantee a usable interface for screen-reader users; unstructured questions can cause an undue burden on short-term memory, but this burden can be minimized by avoiding stem-and-leaf questionnaire structures and using inferences to reduce wordiness. Additionally, problems screen-reader users experience with navigation can be minimized if the programmed tab sequence follows the natural reading order (top to bottom, left to right). Also, although automated accessibility testing software identifies missing labels for graphics and buttons, tabbing order and readability are best checked with screen-reader software. Finally, it is more efficient to create and maintain one accessible Web survey from the beginning of the design process rather than separate text and graphical versions. If a Web survey is accessible and usable, then all users will be able to respond to Web surveys with greater accuracy, ease and satisfaction.

## References.

- [1] HiSoftware (Cynthia Says) Web site <http://www.contentquality.com/>
- [2] HiSoftware (AccVerify) Web site <http://www.hisoftware.com/products/accverify.html>
- [3] SSB Bart Group (InFocus) Web site <https://www.ssbbartgroup.com/amp/infocus.php>
- [4] Moss, Trenton (2007). The Problem With Automated Accessibility Testing Tools, retrieved from <http://www.webcredible.co.uk/user-friendly-resources/web-accessibility/automated-tools.shtml> on 4/10/2009
- [5] GW Micro (Window-Eyes) Web site <http://www.gwmicro.com/>
- [6] Freedom Scientific (JAWS) Web site <http://www.freedomscientific.com/>
- [7] Section 508 (Section 508 Standards) Web site <http://www.section508.gov/index.cfm?FuseAction=Content&ID=12#Software>
- [8] Peterson, L.R., & Peterson, M.J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology*, 58, 193-198.
- [9] Usability.gov (Usability Basics) Web site <http://www.usability.gov/basics>
- [10] JimThatcher.com (Accessible forms) Web site <http://www.jimthatcher.com/webcourse8.htm>

[11] WebAIM (Introduction to Web Accessibility) Web site <http://www.webaim.org/intro/>

Posted in Helpful Ideas, Methods [Comments: 0](#)

[Comments \(0\)](#)

## **Summaries of Address-Based Sampling Presentations at the AAPOR Annual Meeting**

Monday, June 29, 2009, 7:10:51 AM | Editor

### **Building a New Foundation: Transitioning to Address Based Sampling after Nearly 30 Years of RDD**

Michael W. Link, Gail Daily, Charles D. Shuttles, L. Tracie Yancey, Anh Thu Burks, and H. Christine Bourquin, The Nielsen Company

Address based sampling (ABS), the use of a comprehensive database of addresses for sampling of residential households, is garnering considerable attention by survey researchers as a potential alternative to random digit dialing (RDD) surveys. For nearly 30 years, the Nielsen TV Ratings Diary Survey was one of the world's largest RDD surveys (in recent years screening more than 5 million telephone numbers annually). In November 2008, the TV Ratings Diary moved from a landline telephone frame to an ABS frame, becoming the first major survey research effort to make this important transition. The new TV Ratings design uses ABS with a multi mode data collection approach, which includes Web, mail, and telephone recruitment tools. We assess the success of this transition by comparing the March 2009 ABS measurement to the February 2008 RDD design. The lessons learned from this pioneering effort will further the understanding the industry has for the potential uses of this new approach. Some of the key findings included:

Allows researchers to reach cell phone only households

Significant step in improving representation of younger homes

Improves coverage but not necessarily response rate in all cases

Key sample indicators on ABS file are more accurate than corresponding indicators on landline frame — with addition of geocoded information can be a powerful tool for addressing racial/ethnic imbalances

**[Link to Paper](#)**

---

## Using the U.S. Postal Delivery Sequence File for Mixed-Mode Studies: Report on Measurement Differences Between Mail and Telephone Responses in the SHAPE Study

Todd Rockwood, Melissa Constantine , Michael Davern, University of Minnesota  
Timothy Beebe, Mayo Clinic  
Sheldon Swaney, Hennepin County Minnesota

The SHAPE study is a large general population (n=7500) public health screening survey conducted in Hennepin County Minnesota by the Hennepin County Department of Health every three years (<http://tinyurl.com/6w9m3o>). The 2006 administration of the SHAPE study utilized the DSF as the primary sampling frame. Additionally, a mode of administration experiment was conducted in which a sub-sample of respondents were randomized to either the mail (n=1848, RR 70%) or telephone (n=560, RR 27%) mode of administration.

Socio-demographic characteristics

Differences:

- Higher percentage of females (65%) by mail than phone (58%,  $X^2p = .01$ )
- Higher percentage born in the US (93%) by mail than phone (89%,  $X^2p = .01$ )
- Higher percentage live in urban core (as opposed to suburbs, 45%) by mail than phone (40%,  $X^2p = .01$ )

No Differences:

- Race/Ethnicity
- Educational status
- Age

Questions by topical area:

General Health Status – 11 items/ 3 demonstrate significant differences

Depression – 9 items/ 7 demonstrate significant differences

Insurance/Health Services – 6 items/ 4 demonstrate significant differences

Health Behaviors/Screening Tests – 3 items/ 3 demonstrate significant differences

Evaluation of community – 10 items/ 7 demonstrate significant differences

Public/Medical Assistance (Welfare) — 11 items/ 2 demonstrate significant differences

Discrimination – 10 items/ 4 demonstrate significant differences

Overall, 50% of the topical items evaluated demonstrate significant differences between the mail and telephone modes of administration. In general the differences are consistent with what would be expected based on models of mode of administration effects.

---

## Using Address-based Sampling to Survey the General Public by Mail vs. Web plus Mail

Benjamin L. Messer, Don A. Dillman, Washington State University

Our purpose in this study was to determine the extent to which households in an address-based sample would respond to a general public survey via the Internet when the survey request was sent by postal mail. We obtained a statewide random sample of addresses from the USPS Delivery Sequence File (DSF) to conduct the Washington Community Survey (WCS) in the summer and fall of 2008 using both mail and Internet survey modes. Nine experimental treatment groups were designed to test the overall differences between and the effects of different contact procedures, including a \$5 incentive and Internet instruction card, on mail and Internet response. The groups were comprised of a mail-only treatment, three mail preference treatments, and five Internet preference treatments. Respondents in the “preference” groups were mailed a request to complete the survey via a designated mode, mail or Internet, and three weeks later non-respondents were asked to respond by the alternate mode. Among the results:

The \$5 mail only group achieved the highest overall response rate of 56.7%, while the mail-preference groups with a \$5 incentive were close behind, ranging from 53.6% to 55%. Response rates for the Internet groups with a \$5 incentive were significantly lower, ranging from 42.8% to 46.3%. However, offering the Internet as the first mode, as in the Internet preference groups, resulted in two-thirds responding via the Internet and the remaining one-third via postal mail. Offering mail as the first survey mode resulted in very few Internet responses, ranging from only 2% to 5.7% of the total.

The illustrated Internet card, which provided instructions and encouragement to respond via the Internet, did not increase response rates for the two Internet groups that received it. In addition, we found very few significant differences between respondents who received an Internet card vs. those who did not; the only differences were on gender, education, and marital status.

The \$5 cash incentive significantly increased response rates for both mail (52.5% vs. 39.2%) and Internet respondents (31.3% vs. 13.4%) but had a larger impact on Internet response with a difference of 17.9% for the Internet and only 13.3% for the mail. Respondents to the \$5 incentive groups were very similar to respondents who did not receive the \$5, regardless of survey mode; there was a significant difference only for income among mail respondents and on education among Internet respondents.

Respondents to the Internet in the Internet preference groups were, on average, younger, more educated, married, and employed with fewer people in the household and higher incomes

compared to the mail follow-up respondents in these groups. The differences persisted but were much smaller when comparing the Internet respondents to the mail respondents in the mail preference groups. However, these differences become negligible when Internet and mail follow-up respondents in the Internet preference groups were combined and compared to mail respondents in the mail preference groups.

Finally, neither WCS Internet nor mail respondents were completely representative of the Washington population when compared with 2007 American Community Survey (ACS) data for Washington, even with weighting on gender and age. The mail respondents were closer overall but both types of respondents were more educated and more likely to be employed, have children in the household, larger household sizes, and higher incomes compared to ACS respondents. All WCS respondents were representative of the cell-only population in Washington, estimated at about 18%, but were over-representative of households with Internet access by 12% to 24%.

---

### Comparing Random Digit Dial (RDD) and United States Postal Service (USPS) Address-Based Sample Designs for a General Population Survey: The 2008 Massachusetts Health Insurance Survey

Susan Sherr, David Dutwin, SSRS  
Timothy Triplett, Doug Wissoker, Sharon Long, Urban Institute

In the summer of 2008, the Urban Institute and Social Science Research Solutions (SSRS) conducted the Massachusetts Health Insurance Survey (HIS) on behalf of The Massachusetts Division of Health Care Finance and Policy. The goal of the Massachusetts HIS is to document health insurance coverage and access among Massachusetts residents. In an effort to include cell-phone only households in the study, the 2008 HIS employed a dual-sample-frame design that combined a random-digit-dial (RDD) telephone sample and an address-based (AB) household sample. In addition, the AB sample and the RDD sample were each divided into two strata: (1) sample records with both an address and matching telephone number and (2) sample records with either a phone number (RDD) or an address (ABS), but not both. Survey respondents could choose to complete the survey by telephone, web or mail. A total of 4,910 interviews were completed across both sample frames. In comparing results from the RDD and ABS samples, we found that:

The sample yield was better and more efficient in the ABS sample. As a result, the cost per interview was lower with the ABS sample.

Almost half of all respondents—52% of ABS respondents and 34% of RDD respondents—completed the survey online. The popularity of a web questionnaire helped keep **overall survey** costs low.

The ABS response rate (34.7%) was lower than RDD (42.0%). However, there was no significant difference by sample frame or by mode in either breakoffs or incompletes.

8.5% of ABS respondents were from cell-phone only households, a figure that is close to the recent NIH estimate for Massachusetts.

Unweighted demographic characteristics of the ABS respondents were closer to American Community Survey population counts than was true for RDD respondents, indicating a reduction in coverage bias with the ABS. Nevertheless, nonresponse bias among underrepresented groups was present in both sample frames.

We conclude that the ABS offered a number of advantages over the RDD sample, including interviews with cell-only households, better sample yields, and lower costs per interview. The better coverage in the ABS sample results in smaller weights, which resulted in smaller design effects and less sensitivity in estimates of key variables due to weighting as compared to the RDD sample.

[Link to presentation slides](#)

---

## **Performance Rates of CPO Subsequent Survey Households Identified Via Address Frames**

Anna Fleeman, Nicole Wasikowski, Arbitron Inc.

Findings from Arbitron's two large address-based-sample (ABS) studies, fielded in 2007 and 2008, suggest that Cell-Phone-Only (CPO) households can be efficiently identified and less expensively than cellular RDD. A short questionnaire asking about media-related behaviors and cell/landline ownership was sent to the address sample unable to be matched to a phone number (n=20,094). Sampled households responded at encouraging levels with approximately 35% reporting CPO status *and* providing a cell phone number on which to reach them (n=2,058). After constructing the CPO sample pool via the questionnaire mailing, Arbitron then began placing radio listening diaries and encouraging diary return in these CPO households following standard Radio Ratings methodology. Highlights are as follows:

70% of CPO households consented to participate in the diary-based Radio Ratings, nearly double the landline sample (control).

Of the ~2,700 radio-listening diaries sent to consenting households, 72% were returned, which was several points greater than landline sample ( $p \leq .00$ ).

More than double the number of 18-34 year old diarykeepers in CPO sample than landline sample ( $p \leq .00$ ).

Yield of CPO households in ABS greater than cellular RDD: 10% versus 4%.

Using ABS to identify CPO households allows for pre-alert mailings and targeted incentives prior to phone contact.

Because respondents provided their cell phone number on the questionnaire, there is no need to hand-dial as with cellular RDD.

Best of both worlds: maximize mail and phone contact while identifying CPO households.

Using an address frame to include CPO households in Arbitron's Radio Ratings is financially and methodologically better than relying on cellular RDD.

---

### **Modeling the Need for Traditional vs. Commercially-Available Address Listings for In-Person Surveys: Results from a National Validation of Addresses**

Ned English, Colm O'Muircheartaigh, Michael Latterner, Stephanie Eckman, and Katie Dekker, NORC

NORC conducted a national validation of USPS delivery sequence file (or DSF) addresses with three goals: first, to create updated coverage estimates for urban, rural, and suburban areas, as an extension of previous research; second, to learn about DSF coverage in areas that have experienced growth during the past decade, and so have transitioned from rural to suburban or urban; finally, to develop a model to predict DSF coverage based on information available in advance of data collection.

NORC field staff checked the quality of the DSF, confirming addresses that existed, rejecting addresses not present, and adding any new addresses not on the DSF. Based on these results we measured the quality of the DSF in a variety of environments, and determined the relative under- and over- coverage. Our results were as follows:

Urban areas generally have better DSF coverage than rural areas.

Rural areas do not have universally poor coverage.

Traditional listing is not always necessary in rural areas, as rural areas have started to be better represented on the DSF than previously.

Population density and the percent of addresses city style are effective predictors of coverage.

Over-coverage appears to be more haphazard than under-coverage, as the former is dependent on geocoding database quality at the local scale.

---

### **Address Based Sampling and Address Matching: Experience from REACH U.S.**

Katie Dekker and Whitney Murphy, NORC at the University of Chicago

The address-based sampling approach relies heavily on accurate matching of addresses to working telephone numbers. After selecting a sample of addresses, NORC matched selected addresses to telephone numbers using commercial vendors. During the screener portion of the telephone interview, respondents were asked to verify that they live at the selected address. The results of the screening process allow us to assess how well our vendor is able to match addresses and help us to determine whether we can safely eliminate the screener question in future rounds.

At the time of this analysis, our vendor was able to match approximately 70% of the sampled addresses to telephone numbers.

Of the phone numbers that were matched, about 91% of the respondents confirmed the address in the phone interview.

Based on our findings, we do not have the confidence in our match rates to eliminate this screener verification question.

[Link to presentation slides](#)

---

## **Evaluation of Address Based Sampling (ABS) Frame Supplementation Methods for In-Person Household Surveys**

Joseph P. McMichael, Jamie L. Ridenhour, Bonnie E. Shook-Sa and Vincent G. Iannacchione, RTI International

Survey researchers are increasingly looking to Address Based Sampling (ABS) as a less costly alternative to field enumerated sampling frames. Although research suggests that the national household coverage of an ABS frame is high, coverage is not evenly distributed leading to a disparity in coverage between rural and urban areas. The undercoverage of the ABS frame, particularly in rural areas, can create bias in surveys utilizing only an ABS frame. In-person surveys can use field-implemented supplementation methods to increase coverage and reduce this potential for bias. The 2008 American National Election Survey (ANES) is a national, in-person survey that used a frame supplementation procedure called the Check for Housing Units Missed or CHUM (McMichael et al. 2008). This procedure is a series of protocols to systematically identify dwelling units missing from the frame. Evaluation of the CHUM has shown it is a successful method for improving coverage, though more work can be done to improve the training and monitoring of field staff. Future research on the CHUM will focus on improving training for field staff, evaluating cost, and examining bias reduction.

Reference:

McMichael, Joseph, Jamie Ridenhour, and Bonnie Shook-Sa. 2008. A robust procedure to supplement the coverage of address-based sampling frames for household surveys. *Proceedings of the American Statistical Association, Section on Survey Research Methods*.

---

## **Multi-Mode Surveys Using Address Based Sampling: The Design and Preliminary Experience of REACH U.S. Risk Factor Survey**

Martin Barron, NORC at the University of Chicago

This presentation describes the design of the Racial and Ethnic Approaches to Community Health Across the U.S. Risk Factor Survey (REACH RFS). REACH RFS is one of the first large-scale surveys to employ a multimode ABS approach. REACH RFS is a project sponsored by the Centers for Disease Control and Prevention to measure the performance of 28 community-based programs designed to reduce health disparities among racial and ethnic minority populations. REACH RFS will employ ABS combined with data collection via telephone, mail, and face-to-face interviews. In this presentation, we discuss the ABS design and practical implications of the REACH RFS design. In summary:

REACH RFS considered a number of designs but determined an ABS design to be most appropriate.

In designing our ABS approach, the main priorities were to maximize coverage, interviews via the telephone, and response rates.

The final design calls for attempting to contact households via telephone first. If that fails, attempts will be made by mail and face-to-face.

Though it is still too early to gauge the success of this design, we note that this ABS design requires considerably more field time than a traditional RDD survey. We further note that the multi-mode design can quickly lead to an extremely complex design, particularly if respondents are allowed to switch back and forth between modes. Nevertheless, NORC believe that ABS has significant potential for the REACH Risk Factor Survey.

[Link to presentation slides](#)

Posted in [Methods](#) [Comments: 1](#)

[Comments \(1\)](#)

## **Practical Considerations in Raking Survey Data**

Monday, June 29, 2009, 7:09:32 AM | Editor

Michael P. Battaglia, David Izrael, Abt Associates, Inc.

David C. Hoaglin, Abt Bio-Pharma Solutions, Inc.

Martin R. Frankel, Baruch College, City University of New York

## Introduction

A survey sample may cover segments of the target population in proportions that do not match the proportions of those segments in the population itself. The differences may arise, for example, from sampling fluctuations, from nonresponse, or because the sample design was not able to cover the entire target population. In such situations one can often improve the relation between the sample and the population by adjusting the sampling weights of the cases in the sample so that the marginal totals of the adjusted weights on specified characteristics, referred to as control variables, agree with the corresponding totals for the population. This operation is known as raking ratio estimation (Deming 1943, Kalton 1983), raking, or sample-balancing, and the population totals are usually referred to as control totals. Raking is most often used to reduce biases from nonresponse and noncoverage in sample surveys.

Raking usually proceeds one variable at a time, applying a proportional adjustment to the weights of the cases that belong to the same category of the control variable. The initial design weights in the raking process are often equal to the inverse of the selection probabilities and may have undergone some adjustments for unit nonresponse and noncoverage. The weights from the raking process are used in estimation and analysis.

The adjustment to control totals is sometimes achieved by creating a cross-classification of the categorical control variables (e.g., age categories  $\times$  gender  $\times$  race  $\times$  household-income categories) and then matching the total of the weights in each cell to the control total. This approach, however, can spread the sample thinly over a large number of adjustment cells. It also requires control totals for all cells of the cross-classification. Often this is not feasible (e.g., control totals may be available for age  $\times$  gender  $\times$  race but not when those cells are subdivided by household income).

The use of marginal control totals for single variables (i.e., each margin involves only one control variable) often avoids many of these difficulties. In return, of course, the two-variable (and higher-order) weighted distributions of the sample are not required to mimic those of the population.

The next two sections discuss the raking algorithm and its convergence. Subsequent sections discuss control totals and several issues that arise in practical applications: two-variable margins, raking at the state level in national surveys, maintaining adjustments for nonresponse and noncoverage, surveys that involve screening, and weight trimming.

## Basic Raking Algorithm

The procedure known as raking adjusts a set of data so that its marginal totals match control totals on a specified set of variables. The term “raking” suggests an analogy with the process of smoothing the soil in a garden plot by alternately working it back and forth with a rake in two perpendicular directions.

In a simple 2-variable example the marginal totals in various categories for the two control variables are known from the entire population, but the joint distribution of the two variables is

known only from a sample. In the cross-classification of the sample, arranged in rows and columns, one might begin with the rows, taking each row in turn and multiplying each entry in the row by the ratio of the population total to the weighted sample total for that category, so that the row totals of the adjusted data agree with the population totals for that variable. The weighted column totals of the adjusted data, however, may not yet agree with the population totals for the column variable. Thus the next step, taking each column in turn, multiplies each entry in the column by the ratio of the population total to the current total for that category. Now the weighted column totals of the adjusted data agree with the population totals for that variable, but the new weighted row totals may no longer match the corresponding population totals.

This process continues, alternating between the rows and the columns, and close agreement on both rows and columns is usually achieved after a small number of iterations. The result is a tabulation for the population that reflects the relation of the two control variables in the sample. Raking can also adjust a set of data to control totals on three or more variables. In such situations the control totals often involve single variables, but they may involve two or more variables.

Ideally, one should rake on variables that exhibit an association with the key survey outcome variables and that are related to nonresponse and/or noncoverage. This strategy will reduce bias in the key outcome variables. In practice, other considerations may enter. A variable such as gender may not be strongly related to key outcome variables or to nonresponse, but raking on it may be desirable to preserve the “face validity” of the sample.

## Convergence

Convergence of the raking algorithm has received considerable attention in the statistical literature, especially in the context of iterative proportional fitting for log-linear models (Bishop et al. 1975), where the number of variables is at least 3 and the process begins with a different set of initial values in the fitted table (often 1 in each cell). For raking survey data the iterative raking algorithm generally converges after a small number of iterations, say 3 to 10.

Convergence can, however, sometimes require a large number of iterations. Oh and Scheuren (1978) note that the available convergence proofs make strong assumptions about the cell counts in the cross-classification of the raking variables—that no cells are empty or that some particular combination of nonempty cells is present. They recommend setting up the raking problem in a “sensible” manner to avoid: 1) imposing too many marginal constraints on the sample, 2) defining marginal categories that contain a very small percentage of the sample, and 3) imposing contradictory constraints on the sample.

Our experience indicates that, in general, raking on a large number of variables can slow the convergence process. However, other factors also affect convergence. One is the number of categories of the raking variables. Convergence will typically be slower for raking on 10 variables each with 5 categories than for 10 variables each with only 2 categories. A second factor is the number of sample cases in each category of the raking variables. Convergence may be slow if any categories contain fewer than 1% of the sample cases. A third factor is the size of

the difference between each control total and the corresponding weighted sample total prior to raking. If some differences are large, the number of iterations will typically be higher.

One simple definition of convergence requires that each marginal total of the raked weights be within a specified tolerance of the corresponding control total. Tolerance can be defined in absolute terms (e.g., maximum difference less than 10) or in relative terms (e.g., maximum difference less than 0.1 percentage point). As noted above, in practice, when a number of raking variables are involved, one must check for the possibility that the iterations do not converge (e.g., because of sparseness or some other feature in the full cross-classification of the sample). One can guard against this possibility by also setting an upper limit on the number of iterations (e.g., 75). As elsewhere in data analysis, it is sensible to examine the sample (including its joint distribution with respect to all the raking variables) *before* doing any raking. For example, if the sample contains no cases in a category of one of the raking variables, it will be necessary to revise the set of categories and their control totals (say, by combining categories). We recommend, at a minimum, checking the unweighted percentage of sample cases and the percentage of control cases in each category of each raking variable. Small categories in the sample or in the control totals (say under 2%) are potential candidates for collapsing. This step will also reduce the chance of creating very unequal weights in raking. Category collapsing always needs to be done carefully, and in some instances it may be important to retain a small category in the raking.

### Sources and Choices of Control Totals

Surveys that use demographic and socioeconomic variables for raking must locate a source for the control totals. Examples of sources of control totals available in the United States include the 2000 U.S. Census short-form data, the 2000 U.S. Census long-form data, the 2000 U.S. Census 5-Percent Public Use Microdata Sample (PUMS) files, the annual March Current Population Survey (CPS), U.S. Census Bureau population estimates, the American Community Survey (ACS) published estimates, the ACS 2005–2007 PUMS, and private-sector sources such as Claritas, Inc. The ACS is a rolling sample of housing units consisting of around 1.94 million housing units per year. We have used the 2005–2007 ACS PUMS to develop control totals at the state and sub-state level.

If control totals come from more than one source, it is important to make sure that control totals from all sources add to the same population total. If not, the raking will not converge if one is using a maximum-absolute-difference convergence criterion.

One must also consider how the variables are measured. For example, a telephone survey may ask a single question to obtain household income. The source for the control totals, however, may have an income variable that is constructed from a series of questions about income from several sources (wages, cash-assistance programs, interest, dividends, etc.). One needs to consider carefully whether using income as a raking variable makes sense. If the sample is thought to substantially underrepresent low-income persons, then raking on income may be preferred. If, on the other hand, there is concern that the survey is measuring income very differently from the source of the control totals, then consideration should be given to raking on a proxy variable such as educational attainment or even a dichotomous poverty-status variable.

Control totals usually do not come with a “missing” category. The same variable in the survey may have a nontrivial percentage of cases that fall in a DK or Refused category. In this situation it may be possible to impute for item nonresponse in the survey before the raking takes place. When imputation is not feasible, the following procedure can be used to adjust the control totals. Run a weighted frequency distribution on the raking variable in order to determine the percentage of sample cases that have a missing value (e.g., 4.3%). Allocate 4.3% of the overall control total to a newly created missing category (e.g., 4.3% of 1,500,000 = 64,500). Reapportion the control totals in the other categories so that they add to the reduced control total (1,500,000 – 64,500 = 1,435,500). After raking, the weighted distribution of the sample will agree with the revised control totals and will reflect a 4.3% missing-data rate in weighted frequencies and tabulations.

### **Inclusion of Two-Variable Raking Margins**

Raking can be viewed as analogous to fitting a main-effects-only model. Because of sample size limitations and/or availability of only one-variable (factor or dimension) control totals, many raking applications follow this approach. In some situations it may be important to fit a two-variable interaction to the data. For example, one is planning to rake on Variables A, B, C, and D. However, control totals for Variable C crossed with Variable D are available and exhibit a strong interaction (e.g., persons aged 0–17 years are more likely to be Hispanic than persons aged 65+ years). If the cell counts in the  $C \times D$  margin of the sample are large enough to support fitting a  $C \times D$  interaction, one would rake on three margins: A, B, and  $C \times D$ . It is not necessary also to rake on separate margins for Variables C and D. If, however, the  $C \times D$  raking margin involved collapsing, one could consider also raking on one-variable margins for Variables C and D without any collapsing of their categories.

### **Raking at the State Level in a Large National Survey**

Some large national surveys stratify by state and are designed to yield state estimates. The resulting total national sample is usually very large. The survey analysts seek to provide national estimates as well as state estimates. Often one sets up raking control totals at the state level and carries out 51 individual rakings. Assume those rakings use Variables A, B, and C; but the number of categories of each variable is limited because of the state sample sizes.

For example, one might collapse Variables A, B, and C differently by state. If Variable A were race/ethnicity, one might be able to use Hispanic as a separate category in California, but not in Vermont because of the small sample size. After the 51 rakings one might compare the weighted distributions of Variables A, B, and C with national control totals and observe some differences that are caused by the state-level collapsing of categories.

If having precise weighted distributions at the national level is important for analytic or “face validity” reasons, one can use the following raking technique. Set up a single raking that includes margins for  $\text{State} \times A$ ,  $\text{State} \times B$ , and  $\text{State} \times C$  (i.e., combine the 51 individual state rakings into a single raking). Then add detailed national margins for Variables A, B, and C.

Another, similar example would involve adding Variable D as a national raking margin because its control total is available only at the national level (e.g., household income).

This type of raking can also be applied to a state sample that has been stratified into sub-state areas.

### **Maintaining Prior Nonresponse and Noncoverage Adjustments in the Final Weights**

Frankel et al. (2003) have discussed methods based on data on interruptions in telephone service (of a week or longer in the past 12 months) to compensate for the exclusion of persons in nontelephone households in random-digit-dialing surveys. One typically adjusts the base sampling weights of persons with versus without an interruption in telephone service. The resulting interruption-based weight adjusts for the noncoverage of nontelephone households. If one then rakes the sample on age, sex, and race, the impact of the nontelephone adjustment may be diluted somewhat, even though the raking starts with the interruption-based weight.

In that situation it generally makes sense to create weighted control totals (using the interruption-based weight) from the sample for persons residing in households with versus without an interruption in telephone service. These weighted control totals should be ratio-adjusted so that they have the same sum as the age, sex, and race control totals. For example, if the age, sex, and race margins sum to 180,000,000 persons, then the interruption margin must also sum to 180,000,000. The raking would use the four variables instead of just three and would ensure that the nontelephone adjustment is fully reflected in the final weights. This approach would be appropriate where the interruption-in-telephone-service category could be small (e.g., in states where telephone coverage is very high), but one still wants to maintain that small category in the raking.

### **Raking Surveys That Screen for a Specific Target Population**

A common survey model for obtaining interviews with a specific target population is to screen a sample of households for the presence of members of the target population. An example is children with special health care needs. The screening interview collects a roster of children with, say, their age, sex, and race, and determines whether each child has special health care needs. If the household contains one child with special health care needs, a detailed interview is conducted for that child. If the household has two or more such children, one is selected at random for the detailed interview. Of course, the interview response rate will be less than 100%, because some parents will not agree to do the detailed interview.

Assume that, in a national telephone survey, the survey analysts need to look at the prevalence of children with special health care needs, and they will also be analyzing the detailed-interview data. In this situation one would calculate the usual base sampling weights, make adjustments for unit nonresponse, and possibly make a noncoverage adjustment to compensate for the exclusion of children living in nontelephone households if warranted. One first obtains control totals for age, sex, and race in the U.S. population aged 0–17 years. One then rakes the entire sample of children in the screened households to those control totals, because that sample is a

sample of children aged 0–17 in the U.S. The resulting screener weights can then be used to estimate the prevalence of children with special health care needs in the U.S.

That screener weight would typically serve as the input weight in the calculation of weights for the children with completed detailed interviews. As part of that calculation process one also seeks to weight the detailed-interview sample by age, sex, and race. Of course, control totals are unlikely to be available for children with special health care needs. One can, however, use the screener weight and the sample of children with special health care needs identified in the screened households to form weighted control totals for age, sex, and race and then use those totals in raking the detailed-interview weights. This method ensures that the age distribution of children with special health care needs from the screener sample will agree exactly with the distribution in the detailed-interview data.

### **The Original IHB Raking Macro**

Izrael et al. (2000) introduced a SAS macro for raking (sometimes referred to as the IHB raking macro) that combines simplicity and versatility. The IHB raking macro was enhanced to increase its utility and convergence diagnostics (Izrael et al. 2004). The IHB SAS macro produces diagnostic output that contains the following information: iteration number, name of variable currently being raked on, and marginal control total and calculated total weight for each level of the current raking variable, along with their difference and also percentage difference. At termination, the macro gives the iteration number at which termination occurred and the reason, which is either that the tolerance was met or that the process did not converge. The macro also writes diagnostics into the SAS LOG, from several of the checks that it makes.

### **Exhibit 1**

Exhibit 1 illustrates the use of the macro with an example involving two raking variables, VARIABLE1 (containing 3 categories) and VARIABLE2 (containing two categories). The IHB raking macro output shows the weighted distribution of each of the two variables at each iteration (Calculated Margin and Calculated % columns), the corresponding marginal control totals (Marginal Control Total and Marginal Control % columns), and the differences between each control total and weighted sample total (Difference and Difference in % columns). At the start of the raking the sum of the design weights equals 44,850.82, compared to the control total of 62,800.30. After adjusting the design weights so that the sum of the weights in each of the 3 categories of VARIABLE1 equals the corresponding control total, the sum of the weights now equals 62,800.30. Iteration 1 involves first adjusting the weights so that the control totals for the 3 categories of VARIABLE1 (22,154.39, 16,533.88, and 24,112.03, respectively) are satisfied and then adjusting the resulting weights so that the control totals for the two categories of VARIABLE2 (30,697.33, and 32,102.97, respectively) are met. With the convergence tolerance set to a difference of 1, the raking converged after 3 iterations.

Because Variable 2 is the last variable adjusted, there is exact agreement with its control totals. As general guidance we recommend making the most important control variable the last variable in the iteration.

In Exhibit 1 the Calculated % column shows the weighted percentage distribution of the sample, the Marginal Control % shows the percentage distribution of the control totals, and the Difference in % column shows the difference between the two preceding columns.

## Weight Trimming

One limitation of the original IHB macro for raking is that it does not place any limits on the highest and lowest weight values. In some situations the raking may converge, but the resulting weights exhibit considerable variability, as measured by the ratio of the highest weight to lowest weight and by the design effect due to weighting ( $1+cv^2$ , where  $cv$  is the coefficient of variation of the weights). Weight trimming increases extremely low weights and decreases extremely high weights to reduce their impact on the variance of the estimates, especially for subgroup estimates. For example, all weights that are less than  $L$  are increased to  $L$ , and all weights that are greater than  $U$  are reduced to  $U$ . One consequence of the trimming of low and high weight values is that the weights of the entire sample will not add to the population size. Although weight trimming is a separate topic from raking, they are certainly related in the sense that weight trimming typically takes place at the last step in the weight calculations, which is often raking. The objective of weight trimming is to reduce the mean squared error (MSE) of the key outcome estimates. Trimming low and high weight values generally lowers sampling variability but may incur some bias. The MSE will be lower if the reduction in variance offsets the increase in bias. We developed two alternative weight-trimming methods. Both are implemented during the raking process in order to ensure that: 1) limits are placed on low and high weight values in the final weights, and 2) the convergence criteria are satisfied, and the weights sum to the population total. The 2009 SAS Global Forum paper by Izrael, Battaglia and Frankel on weight trimming can be found at:

[http://www.abtassociates.com/attachments/Extreme\\_Survey\\_Weight\\_Adjustment\\_as\\_a\\_Component\\_of\\_Sample\\_Balancing\\_\(a.k.a.Raking\).pdf](http://www.abtassociates.com/attachments/Extreme_Survey_Weight_Adjustment_as_a_Component_of_Sample_Balancing_(a.k.a.Raking).pdf)

## Summary

Raking is often the last weighting step in a survey before producing estimates and analyses. It is sometimes relied on as a “black box” that will improve the quality of the sample by reducing nonresponse bias. We have given some background on how raking works, discussed the convergence process, and indicated items that should be checked before and after the raking. Brick et al. (2003) also discuss issues that one should be aware of when using raking. Estimating standard errors is more complicated if raking has been used to develop the final weights. Readers interested in variance-estimation issues related to raking can consult Deville and Sarndal (1992) and Brick et al. (2000).

The new SAS macro is available for free at:

<http://www.abtassociates.com/Preview.cfm?PageID=40858&FamilyID=8600>

## References

Bishop YMM, Fienberg SE, and Holland PW. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.

Brick JM, Morganstein D, and Valliant R. (2000). Analysis of Complex Survey Data Using Replication. <http://www.westat.com/wesvar/techpapers/ACS-Replication.pdf>

Brick JM, Montaquila J, and Roth S. (2003). Identifying Problems with Raking Estimators. *2003 Proceedings of the Annual Meeting of the American Statistical Association* [CD-ROM], Alexandria, VA: American Statistical Association, pp. 710-717.

Deming WE. (1943). *Statistical Adjustment of Data*. New York: Wiley.

Deville JC, and Särndal CE. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, Volume 87, pp. 376-382.

Frankel MR, Srinath KP, Hoaglin DC, Battaglia MP, Smith PJ, Wright RA, and Khare M. (2003). Adjustments for non-telephone bias in random-digit-dialling surveys. *Statistics in Medicine*, Volume 22, pp. 1611-1626.

Izrael D, Hoaglin DC, and Battaglia MP. (2000). A SAS Macro for Balancing a Weighted Sample. *Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference*, Cary, NC: SAS Institute Inc., pp. 1350-1355.

Izrael D, Hoaglin DC, and Battaglia MP. (2004). To Rake or Not To Rake Is Not the Question Anymore with the Enhanced Raking Macro. *Proceedings of the Twenty-Ninth Annual SAS Users Group International Conference*, Cary, NC: SAS Institute Inc., Paper 207.

Kalton G. (1983). *Compensating for Missing Survey Data*. Survey Research Center, Institute for Social Research, University of Michigan.

Oh HL, and Scheuren F. (1978). Some Unresolved Application Issues in Raking Ratio Estimation. *1978 Proceedings of the Section on Survey Research Methods*, Washington, DC: American Statistical Association, pp. 723-728.

Posted in [Methods](#) [Comments: 0](#)

[Comments \(0\)](#)

## **[An Experimental Test of a Strategy to Maintain Contact with Families Between Waves of a Panel Study](#)**

Monday, June 29, 2009, 7:07:50 AM | Editor

Katherine McGonagle, Mick Couper, and Robert Schoeni, University of Michigan

Keeping track of sample persons between waves of data collection helps minimize attrition in longitudinal studies. All things being equal, the longer the time between data collection waves, the greater the likelihood that sample persons have moved, and the greater the difficulty in locating movers (Couper & Ofstedal, 2007; Duncan & Kalton, 1987). In 1997, the Panel Study of Income Dynamics (PSID) changed from annual to biennial interviewing. To capture residential changes between waves, PSID began sending families a “contact information update” mailing in the year between data collection waves. Families who update or verify the address and telephone information and return it receive a \$10 post-paid check. About half the families responded to this mailing in recent waves. During 2007, families providing this information needed far less tracking or refusal conversion efforts, and half as many contacts to be interviewed, underscoring the cost effectiveness of the mailing.

Given these advantages, we designed a study before 2009 production interviewing to improve the response rate of the contact update mailing. Families were randomly assigned to the following conditions: \$10 as a pre- versus post-paid incentive, mailing design (traditional versus updated), being sent a study newsletter, and timing and frequency of the mailing (July versus October versus both times). This paper reports on initial findings with regard to response rates to the mailing by these different conditions. Overall, there is no effect of incentive type implying that post-paid incentives are more cost effective. Traditional design performs better than updated design. Families receiving a second mailing have higher response rates than those receiving one mailing. There are some interaction effects with timing-of-mailing, with October-only families having higher response rates when also receiving a prepaid incentive and newsletter. Hypotheses for these findings and next steps for analysis are described.

## Background

Various types of contact strategies to improve panel retention have been used in many studies (Couper & Ofstedal, 2007; Laurie et al., 1999). However, we know of no experimental test of particular strategies. While the issue of contact strategy effectiveness is an understudied one, the survey methods literature can be drawn upon for suggestions of ways to improve response rates. First, under some conditions a prepaid incentive increases cooperation by heightening the salience of the incentive, as well as the respondent’s sense of reciprocity (Singer et al., 1999). Moreover, an updated design of respondent contact materials should increase the salience of the request, which may enhance cooperation (Dillman, 2000).

Further, the timing and frequency of the request may affect cooperation. Ideally, the request occurs close enough to the upcoming data collection that most residential changes are captured, but not so close that perceptions of respondent burden are increased, which could occur with too many contacts. PSID interviewing occurs in odd years between March and November. Thus, two times of year were chosen for the mailing: midway between the end of the prior wave and beginning of the next (i.e., July) and as far into the year as was feasible to update addresses before production began (i.e., October). A third condition, July with an October re-mail for non-responders, was chosen to examine the effect on response rates of two contacts versus one.

Finally, evidence on the ideal amount of respondent contact is scant. In the two-year timeline of a biennial survey, respondents participate in a lengthy interview, receive a study newsletter, are

asked to update contact information, and receive a letter alerting them to the upcoming interview. At what point do these multiple contacts become burdensome, or do they in fact enhance perceptions of identification with the survey? It is likely that these perceptions vary by characteristics of sample members. Thus, in this study the manipulation of the mailing involved modifying aspects of the incentive, the design, the timing, and the amount of contact being made, with guidance from evidence in the survey methods literature.

## Methods

### *Conditions:*

PSID families eligible for the 2009 interview (n=8,929) were randomly assigned to four conditions which defined a 2 (“newsletter”) x 3 (“timing”) x 2 (“design”) x 2 (“incentive”) experimental design (Figure 1). To manipulate number of respondent contacts and burden, half the families were sent a study newsletter a year before interviewing began. The second condition was the timing of the mailings with one-third mailed in July, one-third in October, and one-third mailed initially in July with a follow-up mailing in October for nonresponders. Mail design was the third condition, with half the families receiving the traditional black and white design, and half an updated design ([Appendix I](#)). The final condition varied whether the \$10 incentive was pre-paid or post-paid.

### **FIGURE 1. Random Assignment to Experimental Conditions**

### *Results:*

Sixty percent of families provided updated or verified contact information. This did not vary by whether the family was sent a newsletter nor by incentive type. There was a significant effect of timing with a higher rate for July-October follow-up (67.2 percent) compared to July-only (57.8 percent) and October-only (55.8 percent) cases. Finally, traditional design had a significantly higher response rate than updated design (62.3 percent versus 58.0 percent).

Traditional design performed significantly better regardless of newsletter or incentive type. There were no differential effects of incentive by newsletter or design type.

Logistic regression models examined effects of newsletter, incentive, and design within each timing condition. Among July-only cases, traditional design performed significantly better than updated design, and there was no effect of incentive or newsletter. Among October-only cases, a significantly higher response rate was associated with traditional design, pre-paid incentive, and being mailed the newsletter. Finally, there were no significant effects of newsletter, incentive or timing among July-October follow-up cases.

## Discussion

What lessons can be learned from this experiment? First, a follow-up mailing for nonresponders is an effective, low cost strategy that may ultimately reduce the need for expensive tracking

during production. This condition yielded response rates between nearly 10 and 12 percentage points higher than the one-time mailing conditions.

Second, and unexpectedly, traditional design performed better. In both designs, the last known contact information was preprinted on a card that respondents folded over, sealed with an attached sticker, and mailed. The updated design also included lengthy instructions describing the necessity of tearing away the card before folding it. Perhaps the additional instructions made returning the updated card seem complicated and discouraged compliance. We will test this in 2010 by removing the “tear step” and instructions and modifying only the color and design.

Third, the pre-paid incentive performed better than post-paid only in the October condition. Perhaps it was comparatively appealing by October as the U.S economy began to unwind. However, overall pre-paid incentives were not cost effective as they did not increase response rates compared to post-paid. The finding that families who were sent the newsletter had higher response rates in October than those who were not may reflect the importance of maintaining contact with study families. October-only families who were not sent the newsletter had no study contact for at least 10 months, with most having no contact for 16 months or longer. These families had a four percent lower response rate than families who received the newsletter. Taken together with the finding of highest response rates for the mailing condition with follow-up suggests that the most promising timing is an initial contact approximately eight months before production starts, with a follow-up three months later if needed.

We will next examine whether these conditions affect operational burden during 2009 production interviewing, including tracking and refusal conversion rates, and contact attempts needed to obtain a final result. This information will help us design an effective strategy for keeping track of panel families.

## References

Couper, M.P., and Ofstedal, M.B. “Keeping in Contact with Mobile Sample Members.” *Methodology of Longitudinal Surveys*. Ed. Lynn, Peter. New York: Wiley, 2009. 183-203.

Dillman, D.A., *Mail and Internet Survey: The Tailored Design Method*. New York: Wiley, 1987.

Duncan, G.J., and Kalton, G., “Issues of Design and Analysis of Surveys across Time.” *International Statistical Review*, 55.1 (2000): 97-117.

Laurie, H., Smith, R., and Scott, L., “[Strategies for Reducing Nonresponse in a Longitudinal Panel Survey.](#)” *Journal of Official Statistics*, 2.12 (1999): 269-282.

Singer, E.S., Gebler, N., Raghunathan, T., Van Hoewyk, J., and McGonagle, K., “The Effect of Incentives on Response Rates in Face-to-Face, Telephone, and Mixed Mode Surveys: Results of a Meta-Analysis.” *Journal of Official Statistics*, 15.2 (1999): 217-230.

Posted in Helpful Ideas, Methods, Uncategorized [Comments: 0](#)

[Comments \(0\)](#)

## **Analyzing the Cost-Effectiveness of Using Return Receipt and Address Corrections in Mail Surveys**

Monday, June 29, 2009, 7:06:40 AM | Editor

Heather L. Stuckey, Penn State Harrisburg  
Neil Malhotra, Stanford University  
Barbara A. Sims, Marian R. Walters, Penn State Harrisburg

Mail surveys remain an effective and popular mode of data collection, given the relatively low and decreasing response rates of Web surveys (Sheehan, 2001) and the unreliability of e-mail addresses as compared to physical addresses (Crawford et al., 2002; Shannon & Bradshaw, 2002). This article addresses the implications of using return receipts/address updates to reduce the costs of self-administered mail surveys at Penn State Harrisburg (PSH). We build upon Dillman's (2000, 1991) widely-used approach of multiple contact surveys to demonstrate the cost-effectiveness and improvement in completion rates using return receipts.

We found that having valid addresses has the potential of reducing costs associated with a mail survey (e.g. copy costs, postage, etc.). However, in determining the efficacy of the return receipt method *a priori*, two features of the survey must be considered: (1) the relative cost between the return receipt and the additional mailings, and (2) the ratio between usable (deliverable) and non-usable (un-deliverable) addresses (see appendix for a model that demonstrates the utility of return receipts). If the ratio of bad-to-good addresses is higher than the cost of the return receipt relative to the cost of additional mailings, then return receipts are cost-effective.

### **Illustrative Study: Mail Survey of PSH Alumni**

We designed a self-administered, anonymous mail survey to determine factors affecting graduate students' abilities to complete a degree program. A mail survey was utilized for this project instead of a web-based survey, in part because the University did not have current e-mail addresses of students who were no longer enrolled. The study population was the 6,430 students enrolled between 1995 and 2005, including both (a) 3,510 alumni and (b) 2,920 students who did not receive degrees and were not enrolled in Spring 2006 when the survey was distributed.

Many of the questions in the survey used four-point Likert scales and related to student issues such as finances, classes, faculty, advising, and personal experiences. Five mailings were implemented, including: (a) an introductory postcard, (b) a survey, (c) a reminder postcard, (d) a postcard to those who did not complete the degree program, a group with a low (4%) response rate compared to those who graduated (22%), and (e) a duplicate survey. Because a major concern was the validity of addresses to which the postcards and surveys would be sent, an

introductory postcard was used to determine which addresses were valid and which were revised. Therefore, the postcard was mailed return receipt (First-Class).

In all, 377 addresses, which represent 6% of the pool of potential respondents, were updated. In addition, 1,137 undeliverable addresses were removed from the database, resulting in a database of 5,293 valid addresses (instead of the original 6,430). Thus, with 1,420 surveys returned, the true completion rate was 26.8% of 5,293, instead of the 22% which would have been assumed (1420/6430) if the undeliverable addresses had been retained for all steps of the survey.

Because there is limited information on the costs/benefits associated with sending an introductory postcard to collect undeliverable addresses, an important outcome of this project was to assess the financial implications of this approach. There were 1,514 returned postcards, of which 377 included updated addresses. As shown in Table 1, using the corrected addresses collected from the introductory postcard provided a survey cost savings of \$516.40 because mailings subsequent to the introductory card were sent only to valid addresses. This cost savings represents a substantial 5.4% of the full cost of the traditional mail survey before address correction. These figures also understate the benefits of return receipt, because the 377 updated addresses do not lead to cost savings in this analysis, but do enhance completion rates. Since these addresses would not be updated in a traditional mail survey, the “wasted” mailings constitute \$581.89, meaning that the true cost savings of using return receipt in this protocol are \$1,108.29. The \$1,100 represents a cost savings of 11.6% of the full cost of a traditional mail survey pre-correction.

Self-administered mail surveys of larger populations may see a savings in cost by using the return receipt method with the initial postcard mailing. Mail surveys (such as the one described above) which have a large number of “bad” addresses and several additional mailings may benefit from using the return receipt method (see appendix for specifics).

**Table 1. Costs of Survey**

	Cost of Postage and Printing	Traditional Mail Survey (N=6,430)	Mail Survey with “Return Receipt” <sup>1</sup>
<b>Intro Postcard</b>	\$0.12/\$0.31 <sup>2</sup>	\$771.60	\$1,993.30
<b>Survey 1</b>	0.47	3022.10	2487.71
<b>First Reminder Postcard</b>	0.24	1543.20	1270.32
<b>Second Reminder Postcard<sup>3</sup></b>	0.37	1078.92	705.22
<b>Survey 2</b>	0.49	3151.70	2594.57
<b>Total Cost</b>		\$9,567.52	\$9,051.12
<b>Completion Rate</b>		22.1%	26.8%
<b>Wasted Cost<sup>4</sup></b>		\$527.14	\$0.00
<b>Additional Completes</b>		0	101

## Discussion

The current study demonstrates that, when conducting a mail survey with multiple, expensive contacts, an additional element to consider in the design methodology is sending an introductory return receipt postcard for the purpose of identifying invalid addresses. Return receipts are also an efficient means of improving completion rates in mail surveys. By correcting undeliverable addresses, researchers can boost the number of completes rather inexpensively, at least as compared to costly techniques used in RDD telephone interviewing (e.g. callbacks, refusal conversions, monetary incentives).

As the data from the graduate student survey research project at PSH demonstrated, the return receipt postcard strategy provided valuable information regarding relocation patterns of graduate students and changes in addresses. In turn, this effort yielded a substantial savings in overall costs associated with the present study, as well as providing the Alumni Office with updated addresses, thus producing a more valid population list. However, using return receipt may not be cost effective in all cases. Rather, when follow-up mailings are expensive, the subject population is mobile, or the contact list is unreliable, using return receipts can produce cost-savings for researchers.

## References

Crawford, Scott, McCabe, Sean, Couper, Mick, and Boyd, Carol. 2002. "From mail to web: Improving response rates and data collection efficiencies." Paper presented at the International Conference on Improving Surveys, Copenhagen, Denmark.

Dillman, Don A. 2000. "Mail and internet surveys: The tailored design method." New York: John Wiley and Sons, Inc.

Dillman, Don A. 1991. "The design and administration of mail surveys." *Annual Review of Sociology*, 17, 225-249.

Shannon, David M., and Bradshaw, Carol C. 2002. A comparison of response rate, response time, and costs of mail and electronic surveys. *The Journal of Experimental Education*, 70(2), 179-192.

## Acknowledgements

This research was supported in part by a grant from the National Science Foundation ADVANCE Leadership Award (MRW).

## Appendix

The formulas below demonstrate the utility of return using a simple operations research model.

Let  $G$  = the number of "good" (deliverable) addresses

Let  $B$  = the number of “bad” (undeliverable) addresses

Let  $C$  = the cost of sending out mail beyond the introductory postcard

Let  $R$  = the cost of the return receipt

The return receipt will only be cost-effective if:

$$C(G + B) > CG + R(G + B) \text{ [1]}$$

The left-hand side is the cost of sending out additional mail to both “good” and “bad” addresses, but not paying for return receipts. The right-hand side is the cost of only sending out additional mail to “good” addresses, but paying for return receipts for all addresses. Rearranging terms in equation [1], we get:

$$B/G > R/(C - R) \text{ [2]}$$

In other words, by equation [2], return receipts are only cost effective if the ratio of bad-to-good addresses exceeds the relative cost between the return receipt and the additional mailings. As  $C \gg R$ ,  $B/G$  must go to infinity for the inequality to hold, meaning that there is virtually no utility to using return receipts if their cost is about the same as the additional mailings.

Thus, this simple model demonstrates that it is only useful from a cost perspective to use return receipts when the mail survey has two properties: (1) there is a high ratio of bad-to-good addresses; and (2) the cost of the additional mailings is substantial relative to the cost of the return receipt.

In our example,  $B/G = .215$  and  $R/(C - R) = .138$ , so the return receipt is effective since the first number is larger than the second. In other words, the ratio of bad-to-good addresses ( $B/G$ ) is higher than the cost of the return receipt relative to the cost of additional mailings ( $R/(C - R)$ ). The bad-to-good ratio is particularly large (perhaps because the population is mobile) and the cost of the additional four mailings is high (although in line with Dillman’s model of survey implementation). Hence, mail surveys with these features may benefit from using the return receipt method.

Posted in [Helpful Ideas](#), [Methods](#), [Uncategorized](#) [Comments: 0](#)

[Comments \(0\)](#)

## **[Obtaining Responses by Mail or Web: Response Rates and Data Consequences](#)**

Monday, June 29, 2009, 7:04:59 AM | Editor

Glenn D. Israel, University of Florida

This purpose of this study was to explore the willingness of the people who have obtained information from Cooperative Extension, a quasi-general public population, to respond to a customer satisfaction survey via the Internet when receiving the request by postal mail. A list of Extension clients was sampled and randomly assigned to three experimental treatments – the traditional mail-only treatment, a mail/Web choice treatment where the client received both a paper version and the url address to the survey, and a Web-preference treatment. For the latter, the client received only the url in the initial request and then a paper survey and the url were provided in the follow up request.

Results show that Web-preference respondents (who could respond by Web initially and then offered a paper survey later) had a lower response rate (52.6%) than did mail/Web choice and mail-only respondents (59.2% and 64.5%, respectively).

The results also show that respondents in the Web-preference treatment differed by mode. Those who responded by Web were more likely to have used Extension's Solutions for Your Life Web\* site during the past year (36.2%) than did those who responded by mail (7.8%). Web respondents also had an average of 4 more contacts with Extension during the year than did mail respondents (8.6 and 4.1 contacts, respectively).

A larger proportion of Web respondents had at least some college education (85.3%) and were younger (mean age of 53.9 years) than mail respondents (67.7% and 59.8 years, respectively). Perceptions of service quality by Extension's clients varied somewhat by mode of response with more Web respondents being "very satisfied" (76.2%) than were mail respondents (62.1%). The results show that survey mode resulted in differences in the demographic profile of respondents and in the substantive results.

When Web and mail respondents in the Web preference treatment were combined, these respondents were not significantly different with regard to demographic attributes or measures of satisfaction from those in the standard mail-only and mail/Web choice groups.

The finding of differences between those who responded via the Web and those responding by mail in the Web preference treatment should warn survey professionals to avoid relying on the Web alone to conduct surveys of the general public. The similarity of the results, however, between the mail only, mail/Web choice and Web preference treatments suggests that mixed-mode surveys can be considered as an option for collecting data from address-based and client lists. On the other hand, response rates were reduced for the two treatments involving the Internet, more so when clients were pushed to use the Web (as in the case of the Web preference treatment). Given this, researchers might need to increase the initial sample size for mixed-mode surveys to avoid problems of insufficient completes for the data analysis.

\* The *Solutions for Your Life* Website, <http://solutionsforyourlife.ufl.edu/>, is the Florida Cooperative Extension's portal to information on a host of topics, including lawns and gardens, agriculture, the environment, family and consumer concerns, disaster preparation, and much more.

Posted in [Methods](#) [Comments: 0](#)

[Comments \(0\)](#)

## **Survey Practice This Month**

Tuesday, April 28, 2009, 12:04:40 PM | Editor

This month, Survey Practice brings back the Ask the Experts column. Aaron Maitland wrote a short review of the recent literature on labeling scale points for attitude questions.

Ineke Stoop presents some evidence from surveys where attempts to increase response were successful. The paper shows that data quality can be improved with targeted efforts to improve response but that simply increasing response does not necessarily improve the data.

In mail surveys, a particular challenge is to create mail materials that are opened by the sample member. In the article by Emily McFarlane and her colleagues, they show that neither using a stamp instead of metered mail, nor using personalized stickers had any effect on response rates.

We also include in this issue, a short summary of changes to Survey Practice that the editors and the Survey Practice Advisory Board are considering.

John Kennedy            Diane O'Rourke  
David Moore            Andy Peytchev

[survprac@indiana.edu](mailto:survprac@indiana.edu)

Posted in [Monthly Summary](#) [Comments: 0](#)

[Comments \(0\)](#)

## **Should I label all scale points or just the end points for attitudinal questions?**

Tuesday, April 28, 2009, 12:03:37 PM | Editor

Aaron Maitland  
National Center for Health Statistics\*

The decision about whether to label all response scale points or just the end points for attitudinal questions can be an important and vexing decision for question designers. This short article will first discuss theoretical and practical considerations that should help guide the decision about how to label response scale points. Next, I will discuss some of the important empirical findings

on this question from the literature. I will also provide some advice about how to evaluate verbal labels.

The amount of clarity that the labels add to the response scale is the most important consideration in the decision to label scale points. The approach to labeling should be the one that most clearly defines the response scale for respondents. One might argue that labeling of all scale points might offer an advantage in this regard. Several authors concede that it is probably more natural for a person to express his or her opinion using words (Fowler 1995; Krosnick and Fabrigar 1997). However, there is inherent ambiguity in the verbal labels that are frequently used with response scales. For example, people might have different interpretations of what it means to “somewhat favor” a public policy. Conversely, one might argue that even though numbers might be more abstract for most respondents, they might also be more accurate. For example, it has been noted that numbers in response scales at least convey the idea of equal intervals between points on a response scale (Krosnick and Fabrigar 1997). However, respondents can also vary in how they interpret numbers. Schwarz et al. (1991) present evidence that respondents interpret 10 point scales from -5 to +5 quite differently than 10 point scales ranging from 1 to 10. They found that negative numbers imply the opposite of something, whereas the low end of a scale with all positive numbers merely implies the absence of something.

There are many other aspects of the research design that might influence the decision to label scale points. For example, the length of the scale will determine the feasibility of labeling all points. It will be much easier to create labels for 5 point scales than it will be for 11 point scales. As Fowler (1995) writes, “it is difficult to think up adjectives for more than 5 or 6 points along most continua.” The mode of the interview also influences whether or not all scale points can be labeled. Generally, the use of telephone interviewing encourages shorter scales so that the respondent does not have to listen to a long list of response options before answering a question. Furthermore, when longer scales are used in telephone surveys it is more common to label only the endpoints of the scale (Dillman, Smith, and Christian 2008). Data collection methodologies that utilize visual modes of communication can more easily incorporate a full set of labels for response scales.

There are particular challenges with using verbal labels in cross-cultural research. Adequately translating fully labeled verbal scales into other languages is extremely difficult and can require considerable resources. For this reason, some (e.g., Fowler 1995) have suggested that numeric scales might be easier to translate and thus better suited for cross-cultural surveys since the researcher would only need to translate anchors at the ends of a scale. However, there is a dearth of evidence that numeric scales create more comparable survey data and cultural factors can also influence the interpretation of the numbers in a scale (Smith 2003).

Several empirical studies (Krosnick and Fabrigar 1997; Alwin 2007; Saris and Gallhofer 2007) have examined the effect of fully labeling scales versus partial labeling of scales on the quality of the resulting data. The general consensus from these studies is that fully labeled scales produce better data quality than partially labeled scales. Krosnick and Fabrigar (1997) came to this conclusion based on a review of several studies that examined the reliability and validity of fully versus partially labeled scales. Alwin (2007) compared the longitudinal reliability of 26 seven-

point scales with only the endpoints labeled with 11 fully labeled seven-point scales from the National Election Studies. This study found that the fully labeled scales had a reliability of .719, whereas the scales with only the endpoints labeled had a reliability of .506. Saris and Gallhofer (2007) also concluded that labeling all points had a positive effect on reliability based on a meta-analysis of 1023 survey questions from 87 multi-trait multi-method experiments. Labeling did not have a significant effect on validity according to their meta-analysis.

There are a couple of other reasons to believe that labeling might lead to better data quality. First, improvements in both reliability and validity tend to be the greatest amongst respondents with lower levels of education (Krosnick and Fabrigar 1997). This is a group that frequently encounters comprehension problems in surveys and question designers are often looking for design strategies to improve data quality among this group. Second, fully labeling scales seem to reduce the effects of question features that are unrelated to the response task. For example, a study by Tourangeau, Couper, and Conrad (2007) conducted experiments using Web surveys that varied the color of the shading on response scales and the use of labels. They found that the effect of color on respondents answers – something designers would not intend – disappeared when the response scales were fully labeled. Hence the literature reviewed in this article seems to suggest that fully labeling has a number of advantages over labeling only the endpoints.

The decision about which labels to use is just as important as the decision about whether or not to use the labels at all. Saris and Gallhofer's (2007) meta-analysis offers some guidelines for selecting labels. Generally, symmetrical labels tend to yield higher reliability and validity. Verbal labels that match the numbers on the scale lead to higher reliability. For example, it might be better to have bipolar scales with negative labels matching negative numbers and positive labels matching positive numbers. Finally, if one does decide to label only the endpoints, these labels should represent fixed reference points (e.g. completely dissatisfied – completely satisfied, instead of dissatisfied-satisfied) so that the end of the scale is well defined.

Even though there are some guidelines, it is always necessary to evaluate a question to determine the best labeling. There are a number of question evaluation methods that might be useful for designing scale labels. Qualitative techniques such as cognitive interviewing that make use of probing or thinkaloud methods will help to understand how respondents assign meaning to the labels on a response scale and whether that meaning matches the survey designer's intended meaning. Item response theory modeling approaches provide a useful quantitative tool to assess whether the pattern of responses matches the theoretical underpinnings of the response scale. For example, one can assess whether the points on the scale are increasingly difficult to endorse as one expresses more extreme attitudes. It is always good practice to use a combination of qualitative and quantitative methodologies such as these to design good survey questions.

In conclusion, the approach to labeling should be the one that most clearly defines the response scale for respondents. The existing empirical literature generally suggests that fully labeled scales have an advantage over partially labeled scales according to a number of criteria. However, practical considerations such as the mode of the interview, number of scale points, and the population under study can play an important role in deciding the extent of labeling that should be used. Additionally, question designers should rely on appropriate question evaluation methods to determine which labeling approach is best for a specific research design.

*\* The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the Centers for Disease Control and Prevention.*

## References

Alwin, D.F. *Margins of Error: A Study of Reliability in Survey Measurement*. New York, NY: John Wiley and Sons, Inc. 2007.

Alwin, D.F., and Krosnick, J.A. "The Reliability of Survey Attitude Measurement: The Influence of Question and Respondent Attributes." *Sociological Methods and Research* 20 (1991): 139-181.

Dillman, D.A., Smith, J.D., and Christian, L.M. *Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. New York, NY: John Wiley and Sons, Inc. 2008.

Fowler, F.J. *Improving Survey Questions: Design and Evaluation*. Thousand Oaks, CA: Sage

Krosnick, J.A. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology* 5 (1991): 201-219.

Krosnick, J.A., and Fabrigar, L.R. "Designing Rating Scales for Effective Measurement in Surveys." *Survey Measurement and Process Quality*. Ed. Lyberg, L., Biemer, P., Collins, M., de Leeuw, E., Dippo, C., Schwarz, N., Trewin, D. New York, NY: John Wiley and Sons, Inc. 1997. 141-164.

Saris, W.E. and Gallhofer, I.N. *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. John Wiley and Sons, Inc. 2007.

Schwarz, N., Knauper, B., Hippler, H.J., Noelle-Neumann, E., and Clark, F. "Rating Scales May Change the Meaning of Scale Labels." *Public Opinion Quarterly* 55 (1991): 618-630.

Smith, T. "Developing Comparable Questions in Cross-National Surveys." *Cross-Cultural Survey Methods*. Harkness, J., Van de Vijver, F., Moher, P. New York, NY: John Wiley and Sons, Inc. 2003. 69-91.

Posted in Ask the Experts [Comments: 5](#)

[Comments \(5\)](#)