Web Survey Design and Measurement Error

Mick P. Couper, PhD
Survey Research Center, University of Michigan
and
Joint Program in Survey Methodology

In celebration of Jelke Bethlehem
Leiden, 19th September, 2014

Introduction

- Over the last 15+ years*, the Internet has had a profound effect on survey data collection
- Rapid and widespread adoption despite continuing debate about inferential issues
- The Web has also had a big effect on research into survey measurement
  - New question design tools and technologies expand the range of what can be done
  - Ability to randomize and relatively cheap data collection facilitate experiments
- I present selected examples to illustrate two points
  - The range of research on Web survey design
  - The importance of design in minimizing measurement error

*CASIC 1998 – see next slide
CASIC 1998

- InterCASIC conference in December 1996
- Marked the emergence of the Internet as a data collection mode
- Brief mention of WWW in one chapter on business surveys
- Brief mention of early tests of online data collection

Selected Examples

- Choice of input tools and response order effects*
- Grouping response options*
- The design of text fields
  - Currency questions*
  - Month and day questions*
  - Asking date of birth
- Use of images
  - Assimilation effects*
  - Contrast effects*
- Interactive features
  - Definitions
  - Running tallies*
  - Visual analog scales
  - Navigating grids or matrixes

*Selected results included
Input Format and Visibility of Options

The Visibility Principle (WTSIWWG)

- Response order effects are well known
  - Primacy effect: items at the beginning of a list are more likely to be endorsed than items at the end of the list
  - Visser et al. (1998): average advantage of 3.1 percentage points for candidate listed first on ballot
  - Believed to be due to satisficing (sub-optimal responding)
  - (Recency effect more common in telephone surveys)
- We went one step further to look at response options that are not initially visible, or require extra actions to make visible
Response Format Experiment

- Couper, Tourangeau, Conrad, & Crawford (2003, *Social Science Computer Review*)
- Randomly assigned to treatment in 2x3 design
  - Order of items:
    - One condition reversed order of response options
  - Format of question:
    - Radio button
    - Drop box, no responses initially visible
    - Drop box, 5 responses initially visible
- About 2,900 respondents from an access panel completed the survey

Radio Button Version A

Which of the following nutrients is most important to you when selecting breakfast cereal? *(Please select one)*

- Protein
- Carbohydrates
- Sugar
- Fat
- Fiber
- Vitamin A
- Vitamin C
- Calcium
- Iron
- Vitamin E
- None of the above
Radio Button Version B

Which of the following nutrients is most important to you when selecting breakfast cereal? (Please select one)

- Vitamin E
- Iron
- Calcium
- Vitamin C
- Vitamin A
- Fiber
- Fat
- Sugar
- Carbohydrates
- Protein
- None of the above

Drop Box, None Visible: Before

Which of the following nutrients is most important to you when selecting breakfast cereal? (Please select one)
Drop Box, None Visible: After

Which of the following nutrients is most important to you when selecting breakfast cereal? (Please select one)

Drop Box, Five Visible, Versions A and B
Cereal: Percent Choosing Protein-Fiber, by Position and Format

<table>
<thead>
<tr>
<th>Position</th>
<th>Format</th>
<th>First 5</th>
<th>Second 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Radio</td>
<td>61.4</td>
<td>50.3</td>
<td>54.8</td>
</tr>
<tr>
<td>Drop none</td>
<td>59.8</td>
<td>54.8</td>
<td>54.8</td>
</tr>
<tr>
<td>Drop 5</td>
<td>67.1</td>
<td>40.0</td>
<td>40.0</td>
</tr>
</tbody>
</table>

Car: Percent Choosing Car Features by Position and Format for Car Item

<table>
<thead>
<tr>
<th>Position</th>
<th>Format</th>
<th>First 5</th>
<th>Second 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Radio</td>
<td>48.3</td>
<td>43.3</td>
<td>40.5</td>
</tr>
<tr>
<td>Drop none</td>
<td>47.4</td>
<td>40.5</td>
<td>40.5</td>
</tr>
<tr>
<td>Drop 5</td>
<td>52.4</td>
<td>40.0</td>
<td>40.0</td>
</tr>
</tbody>
</table>
Conclusions

- Evidence of primacy effects
- Alternative mechanisms:
  - Respondents may just select the first acceptable answer they read, not bothering to read the later options at all
  - Respondents may read all options, but spend less cognitive resources on processing the bottom options compared to the top ones
- We see evidence of both these mechanisms from eye-tracking studies
Grouping Response Options

- When respondents are faced with finding responses in long lists, will organizing the list help?
  - E.g., use of headings, columns or spacing
- Example: highest education question

Grouping Example: ACS Education Item
Analysis of the ACS Education Item

- Horwitz et al. (2012, Census Bureau report)
- Analysis of paradata (mouse clicks) from 19,406 households who accessed the instrument
- This item was the most frequently changed in the entire survey
  - 10.6% of times the answer was changed on this screen before clicking “next”
  - That is, even though one answer was required, respondents were making multiple selections
- We designed an experiment to explore this

Response Grouping Experiment

- Redline, Tourangeau, Couper, Conrad, and Ye (2009, Federal Committee on Statistical Methodology conference)
- Using both paradata and lab-based eye-tracking to explore how respondents navigate through long lists with navigation aids
- Experiment with subjects randomized to alternative versions
  - One column versus three columns
  - Headings versus no headings
Two Items Tested from Paper ACS

What is the highest degree or level of school this person has COMPLETED? Mark (x) ONE box. If currently enrolled, mark the previous grade or highest degree received.

- No schooling completed
- Nursery school to 4th grade
- 5th grade or 6th grade
- 7th grade or 8th grade
- 9th grade
- 10th grade
- 11th grade
- 12th grade, INC. DIPLOMA
- HIGH SCHOOL GRADUATE — high school DIPLOMA or the equivalent (for example: GED)
- Some college credit, but less than 1 year
- 1 or more years of college, no degree
- Associate degree (for example: AA, AS)
- Bachelor’s degree (for example: BA, AB, BS)
- Master’s degree (for example: MA, MS, MEng, Med, MSW, MBA)
- Professional degree (for example: MD, DDS, DVM, D.D.S., D.V.M.)
- Doctorate degree (for example: PhD, EdD)

How did this person usually get to work LAST WEEK? If this person usually used more than one method of transportation during the trip, mark (x) the box of the one used for most of the distance.

- Car, truck, or van
- Bus or trolley bus
- Streetcar or trolley car
- Subway or elevated
- Railroad
- Ferryboat
- Taxi cab
- Motorcycle
- Bicycle
- Walked
- Worked at home — Skip to 27
- Other method

One Column, With Headings
One Column, No Headings

What is the highest degree or level of school you have completed?
Choose ONE only. If currently enrolled, choose the previous grade or highest degree received.
- No schooling completed
- Nursery school to 4th grade
- 5th grade or 6th grade
- 7th grade or 8th grade
- 9th grade
- 10th grade
- 11th grade
- 12th grade, no diploma
- High school graduate - high school diploma or the equivalent
- Some college, no degree
- Associate degree
- Bachelor's degree
- Master's degree
- Professional degree
- Doctoral degree

Three Columns, No Headings

What is the highest degree or level of school you have completed?
Choose ONE only. If currently enrolled, choose the previous grade or highest degree received.
- No schooling completed
- Nursery school to 4th grade
- 5th grade or 6th grade
- 7th grade or 8th grade
- 9th grade
- 10th grade
- 11th grade
- 12th grade, no diploma
- High school graduate - high school diploma or the equivalent
- Some college, no degree
- Associate degree
- Bachelor's degree
- Master's degree
- Professional degree
- Doctoral degree
Field Experiment Results

- Based on client-side paradata
- Percent selecting more than one option (education item); results for transport item similar
- Label and column effects both significant; interaction marginally significant

<table>
<thead>
<tr>
<th>Format</th>
<th>No heading</th>
<th>Heading</th>
</tr>
</thead>
<tbody>
<tr>
<td>One column</td>
<td>7.3% (481)</td>
<td>21.8% (478)</td>
</tr>
<tr>
<td>Three columns</td>
<td>4.4% (483)</td>
<td>19.6% (479)</td>
</tr>
</tbody>
</table>

Percent Making More than One Selection by Headings

χ²(1) = 142.1, p < .0001 for education, χ²(1) = 30.4, p < .0001 for transport
Eye-Tracking Results from Laboratory

- One-column heading condition slows respondents down
- Total fixation time (in ms) across both items, by condition

<table>
<thead>
<tr>
<th>Format</th>
<th>No heading</th>
<th>Heading</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>One column</td>
<td>7.8</td>
<td>14.5</td>
<td>11.2</td>
</tr>
<tr>
<td>Three columns</td>
<td>7.8</td>
<td>9.8</td>
<td>8.7</td>
</tr>
<tr>
<td>Total</td>
<td>7.8</td>
<td>12.5</td>
<td>(n=97)</td>
</tr>
</tbody>
</table>

- Heading main effect significant (F(1,96)=12.5, p<.001);
  column main effect and interaction effect not significant (p>.05)
Implications for Design of Lists

- Results suggest that use of headings may interfere with search in lists
- Headings suggest one selection must be made from each group
- The result is an inefficient process
- The effects would not be detected looking at the substantive responses
Guiding Respondents to Format Answers

- Without guidance, respondents don’t know how to answer open-ended questions
  - Currency amounts
  - Month and year or date questions
  - Other structured formats (e.g., telephone number, zip code, state/province, etc.)
- By providing visual guidance (using templates) or appropriate input tools, we can help them get it right the first time
- Three examples:
  - Templates for currency amounts*
  - Month and year*
  - Date of birth

Experiment on Use of Templates

- Experiment in opt-in panel, surveys conducted in 2006 and 2007
- Crossed a short (6-character) and long (20-character) input field with template or no template, for two dollar amount questions:
  - “How much did you spend last month on prescription drugs?”
  - “How much did you spend last month on non-prescription drugs?”
- We found only a small effect of input field size
- We focus here on the effect of templates
Results of Template Experiment

- Examined entry of non-numeric data:
  - Dollar sign ($)
  - Decimals
  - Other characters
- Multiple types of ill-formed responses are possible (e.g., “about $53.00”)
- Significant differences in proportion of ill-formed responses (p<.0001) for both items
Percent Ill-Formed Entries by Use of Template

Use of Templates Reduces Number of Ill-Formed Entries (Prescription Drug Item)
Month and Year Experiment 1

- Two questions asking for month and year:
  - “In what month and year did you last see a medical doctor?”
  - “In what month and year did you last see a dentist?”
- Three versions of input:
  - Single long input field (30 characters)
  - Two separate fields, one for month, one for year (15 characters each)
  - Drop boxes, one for month (January to December), another for year (last 10 years in descending order, then “Prior to 1995”)
- About 1,100 respondents per condition
Month and Year Experiment 2

- Couper, Kennedy, Conrad, and Tourangeau
- Same two questions:
  - “In what month and year did you last see a medical doctor?”
  - “In what month and year did you last see a dentist?”
- Three versions of input:
  - Two separate fields, labeled “Month” and “Year”
  - Two separate fields, labeled “MM” and “YYYY”
  - Drop boxes, one for month, another for year
- About 800 respondents per condition
Month/Year Experiment 2 Versions

Words:

In what month and year did you last see a medical doctor?
Month:  
Year:  

Next

Symbols:

In what month and year did you last see a medical doctor?

MM:  
YYYY:  

Next

Drop boxes:

In what month and year did you last see a medical doctor?
Month:  
Year:  
Select  
Select  

Next

Responses by Label and Input Type, Last Doctor Visit*

* Similar result for last dentist visit
Conclusions on Input Format

- Question wording is often not enough to shape how respondents answer open questions
- Providing visual guidance reduces formatting errors
  - Reduces the need for error messages and repeat entry
  - Reduces the need for post-survey coding or data cleaning
- Helps both respondents and analysts

Images as Context
Effect of Images

- It is very easy to add images to Web surveys
- Images draw attention
- Images added to questions may have unintended effects on question interpretation
  - Images are necessarily concrete and may affect construal of category membership
  - Images, like prior survey questions, are powerful contextual stimuli
- Examples
  - Assimilation effects
  - Contrast effects

Assimilation Effects of Images

- Couper, Tourangeau, & Kenyon (2004), *Public Opinion Quarterly*
- Use of Knowledge Networks panel
- Sample size of 2,385
- Independent randomization to different pictures for several measures
  - Travel
  - Sporting events
  - Live music events
  - Dining out*
  - Shopping*
Shopping: Clothing Store Version

How many times have you gone shopping since March 1st this year?
Enter an answer from 0 to 100

Shopping: Grocery Store Version

How many times have you gone shopping since March 1st this year?
Enter an answer from 0 to 100
Shopping: Results

- Significant (p<.01) effect on mean number of shopping trips in past month
- No interaction effects with respondent gender

Dining Out: Fancy Restaurant Version

About how many times have you eaten out since March 1st this year?

Enter an answer from 0 to 100
Dining Out: Fast Food Version

About how many times have you eaten out since March 1st this year?
Enter an answer from 0 to 100

Dining Out: Results

- Significant (p<.01) effect on mean number of times dined out in past month
- Similar significant effects for enjoyment of last meal and amount paid for last meal
Summary of Results

- Picture of high frequency activities yielded consistently higher reports of behavior than pictures of low frequency activities
- Effect of pictures carried over to subsequent questions on these activities
- In a subsequent study, we replicated and extended this experiment in the Netherlands

Replication and Extension

- Toepoel and Couper (*Public Opinion Quarterly*, 2011)
- Use of MESS/LISS panel, May and June 2009
- May: 5106 respondents, 64.3% response rate
- Same 5 image questions as earlier study
- Test whether visual effects can be reduced through verbal instructions
Experimental Design

- 3 image conditions
  - No picture
  - Low frequency picture
  - High frequency picture
- 3 instruction conditions:
  - No instruction
  - Instruction to include only low frequency instances
  - Instruction to include both high and low frequency instances
- One condition (high frequency picture, low frequency instruction) has counteracting visual and verbal cues

High Frequency Image, No Instruction
High Frequency Image, High and Low Frequency Instruction

High Frequency Image, Low Frequency Instruction
Example: Frequency of Shopping Events

Summary of Results

- The no-instruction condition replicates earlier findings
- Summary of MANOVA
  - Main effect instructions: lambda=.597, p<.0001
  - Main effect pictures: lambda=.964, p<.0001
  - Interaction instructions*pictures: lambda=.9691, p<.0001
- While both the main effect and interaction are significant, instructions explain more of the variation in the rates than pictures
- Verbal instructions diminish the effects of visual cues
- Respondents use verbal instructions when confronted with counteracting verbal and visual cues
  - May also be order of presentation
Contrast Effect of Images

- Consider two examples of health surveys:

- The images change the meaning of the concept “health”
- These may produce contrast effects


Across 3 experiments, we have varied:

- The content of the image (see next slides)
- The location of the image: question area, header, or section introduction
- The size of the image: small versus large
- The position in the questionnaire: beginning or middle

Convenience samples of opt-in panel members with randomization to condition

Same images used for all experiments

Self-rated health measure: varied format of scale
**Example: Image in Header**

<table>
<thead>
<tr>
<th>Health Images</th>
<th>Health Images</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Health Image 1" /></td>
<td><img src="image2.png" alt="Health Image 2" /></td>
</tr>
</tbody>
</table>

**How would you rate your health?**

<table>
<thead>
<tr>
<th>Extremely good</th>
<th>Good</th>
<th>Neutral</th>
<th>Poor</th>
<th>Extremely poor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Questions about this survey? Email us at [info@research.com](mailto:info@research.com) or call 1-888-123-4567.
Study 1: Percent Reporting Very Good to Excellent Health

- When response scale was: excellent, very good, good, fair, poor
- Significant (p<.05) effect of images on self-reported health
- Significant interaction of picture and location

![Bar Chart]

Study 1: Effect of Image and Location

![Line Graph]

* Scale range of 1-5; a high score indicates good health
Study 2: Percent Reporting Very Good to Excellent Health

- Results for all scale versions, collapsed into Excellent/very good versus other
- Significant (p<.001) main effect of image on self-reported health
- Effect of image is strongest for fully-labeled 5-point scale and weakest for 10-point end-point labeled scale

Study 2: Mean Health Ratings, by Image Content and Position

* Scale range of 1-5; a high score indicates good health
Study 3: Percent Reporting Very Good to Excellent Health

- Response scale 1=very poor to 10=excellent
- Significant (p<.01) main effect of images on self-reported health
- Percent of respondents reporting their health as 8-10 on the scale

Study 3: Mean Health Ratings, by Image Content and Position

* Scale range of 1-10; a high score indicates good health
Summary of Results

- Contrast effect is robust
  - 3 of our studies
  - Replications by others
- The size and placement of the images relative to the question don’t seem to have much effect
- Whether images in the header affect responses depends on whether respondents pay attention to the header

Images and Context: Conclusions

- People attend to images
- Consciously or unconsciously, they can affect answers, altering
  - The construal of the category
  - The specific instances that come to mind
  - The standard to which the target item is compared
- But images may be ignored by some, if they are not clearly linked to the question
- Effect of images may be countered by verbal instructions
Running Tallies or Totals

Running Tallies

- Constant sum or running total questions are common in both household and business surveys
- Can we exploit the interactive features of Web surveys to improve the quality of data?
- Constant sum can be determined by:
  - Designer (e.g., 100%, 24 hours)
  - Respondent, based on previous answer (e.g., how many overnight trips have you taken)
Tally Experiment 1

- Conrad, Couper, Tourangeau, Galesic (2005 AAPOR)
- Task: allocation of percent time on various online activities
- 3 conditions:
  - No feedback/no message
  - Delayed message if submitted tally ≠ 100%
  - Concurrent tally + delayed feedback
- Opt-in panel; about 600 respondents per condition
Thinking of all of the time that you use the internet, what percentage of the time do you spend on the following activities? Please do not count the same activity categories more than once.

Please be sure your answers add up to 100%.

Your answers do not add up to 100%. Please revise your answers so that they add to 100%.

- EMAIL - composing and reading messages
- NEWS - reading newspapers and news magazines; include weather, sports, and financial information
- RETRIEVAL INFORMATION, for example, with a search engine like Google
- INSTANT MESSAGING and CHATTING
- FRANCHISES, buying and selling merchandise, stocks, services, etc.; do not include purchases for travel
- TRAVEL PLANNING - transportation and lodging information, reservations, purchase, getting maps and directions
- VIDEO and MUSIC - downloading or streaming music, videos, movies, etc.; do not include time spent viewing downloaded files
- PLAYING GAMES - with remote players or at games sites; do not include time spent playing games downloaded from a web site
- TAKING A COURSE - distance learning, only include time spent actually on line
- OTHER
- TOTAL

Feedback Improves “Accuracy,” Especially If Concurrent

All differences statistically significant (p<.05)
Tally Experiment 2

- Conrad, Yan, Tourangeau, Couper (AAPOR 2007)
- Task: 24-hour time diary (hours and minutes)
- Opt-in panel; about 400 respondents per condition

### Tally Experiment 2: Time Diary

Section B: Time Use

Your total time does not add to 24 hours. Your current total is 26 hours and 30 minutes. Please update your answers.

We are interested in learning how people balance their time between work, family, and other activities. This question is about how much time you spent on each activity in the last 24 hours.

In the last 24 hours, how much time did you spend on the following activities? Please provide times to nearest minutes.

<table>
<thead>
<tr>
<th>Activity</th>
<th>Hours</th>
<th>Minutes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sleeping</td>
<td>12</td>
<td>00</td>
</tr>
<tr>
<td>Working</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>Taking care of children in the household</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Taking care of other household members</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eating and drinking</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shopping</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Telephone calls, mail and e-mail</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leisure and sports</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Religious activities</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Educational activities</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td>28</td>
<td>30</td>
</tr>
</tbody>
</table>

Next  | Back
Conclusions: Running Tallies

- Providing client-side (concurrent) feedback increases well-formed answers and improves accuracy
- The more complex the task, the more feedback helps
Conclusions From Measurement Research

- It’s not just the words that matter in Web and other survey modes – the visual presentation of questions and response options affect answers.
- Designers must pay attention not only to the semantics (the question wording) but also to the syntax (how the words and visual elements are arranged on the screen and presented to the respondent).
- Web surveys can also give us insights into how people answer survey questions.

Future Challenges for Web Survey Design

- How do we exploit the visual and interactive features of the Web to enhance and extend measurement?
- How do we adapt to mobile Web (smartphones and tablets)?
  - Exploit the mobility aspects of devices (e.g., GPS)
  - Avoid design problems (e.g., small screen)
- How do we best use the Web in a mixed-mode environment?
  - Unimode design versus optimizing for modes?
To Learn More About Our Research

- Search the bibliography at [www.websm.org](http://www.websm.org)

Thank you!

And congratulations to Jelke on your retirement and an exceptional career!

Enjoy the rest of the journey