Questionnaire Design, Development, Evaluation, and Testing: Where in the World Are We?

Gordon Willis, PhD
National Cancer Institute, NIH

QDET2 Keynote Address
November 10, 2016
Miami, Florida
Cognitive Interviews of Questionnaire Designers/Evaluators

Special thanks to:

- Paul Beatty, US Census Bureau
- Jen Dykema, University of Wisconsin
- Jack Fowler, University of Massachusetts
- Emily Geisen, RTI
- Kerry Levin, Westat
- Kristen Miller, NCHS/CDC
Questions I address in my presentation

(I) WHERE ARE WE TODAY?

(II) WHAT DOES THIS MEAN? WHY DOES IT MATTER?

(III) WHERE ARE WE GOING?
• “The world is full of well-meaning people who believe that anyone who can write plain English and has a modicum of common sense can produce a good questionnaire”
  – (A. Oppenheim, 1966)

• Evidence?
  – Many examples of misguided applications
  – Some predating any attempt to develop QDET
QUESTIONNAIRE

1. What should you most like to do, to know, to be? (In case you are not satisfied).
2. Why wouldn’t you change places with any other human being?
3. What do you look forward to?
4. What do you fear most from the future?
5. What has been the happiest moment of your life? The unhappiest? (If you care to tell).
6. What do you consider your weakest characteristics? Your strongest? What do you like most about yourself? Dislike most?
7. What things do you really like? Dislike? (Nature, people, ideas, objects, etc. Answer in a phrase or a page, as you will).
8. What is your attitude toward art today?
9. What is your world view? (Are you a reasonable being in a reasonable scheme?)
10. Why do you go on living?
A short history of questionnaire design models/theory

- Cantril and Fried (1944): Prehistoric cognitive probing
- Payne (1951): Rules – *cookbook* rather than *theory*
- Lansing, Ginsburg, and Braaten (1961), U IL:
  - *An Investigation of Response Error*
- Cannell, Miller, Oksenberg (1981), U Michigan
The Four Horsemen of the Questionnaire Design Apocalypse (Tourangeau, 1984)

- Comprehension
- Retrieval
- Decision/Judgment
- Response Matching
Cognition and Survey Methodology (CASM) as a catalyst for change

• CASM: Interdiscipline between cognitive psychology, survey methods

• Cognitive Laboratories (1980’s):
  ○ Monroe Sirken, NCHS Cognitive Lab
    Census Bureau, BLS...
    Applied Lab as a tangible enterprise:
  ○ Took advantage of adaptation of *Think-Aloud* (Ericsson-Simon) as advocated by Elizabeth Loftus, and *Verbal Probing*
Beyond HOW to ask: WHAT to ask- *are we asking the right Q* in the first place?

This notion has persisted, under different guises:

– Classically, as *Specification Error* – Lazarsfeld (1986) -- we want one thing, but measure another

– Recognition that 4-stage cognitive model didn’t capture all problem types:
  
  o Willis, Royston, Bercini (1991): Added *Logical/Structural errors*
  
  o Problems are more ‘in the question’ than ‘in the respondent’

-- So, we need a fifth horseman ->>>
The Four+1 Horsemen of the Questionnaire Design Apocalypse

Comprehension  Retrieval  Decision/Judgment  Response Matching

Logical/Structural
– The logical extension of CASM arguing that “there’s something else out there” – linguistics, sociology, anthropology...

• Eleanor Gerber, Census; Kristen Miller, NCHS
• Royston: Little of what we do is really ‘cognitive’; Gerber and Wellens (1997) – “Cognition” in the Cognitive Interview?
• Schaeffer and Dykema (2011): Cognitive versus Interactional models
• Multi-interdisciplinary view was pronounced by time of CASM II in 1997
• Landscape shifted: CASM III → MIST->QUEST, QDET, QEM
• Focus is on methods – how much is ‘theory’ involved?
  – Dillman: Our methods are dependent variables...
From Lite to Saturated model*

*Inspired by Norm Bradburn, QDET1
From Lite to Saturated model*

*Dillman: Shift from IAQ to SAQ-Web – Friday Invited Presentation

*Inspired by Norm Bradburn, QDET1
From Lite to Saturated model

Linguistic attributes
Logical/Structural features

Background:
Sociocultural/linguistic context
From Lite to Saturated model

Characteristics
Interaction-relevant behaviors

Investigator/Designer → Survey Question → Respondent

Background:
Sociocultural/linguistic context
From Lite to Saturated model

Motivation
Perception
Cognition
Demographics
Personality
How do we deal with all this complexity?

Characteristics

Logical/Structural features

Motivation

Perception

Cognition

Demographics

Personality

Interaction - relevant behaviors
Different methods target different subsets

Cognitive testing
Different methods target different subsets

Jans: Presentation on Behavior Coding system-- Saturday

Behavior Coding
Different methods target different subsets

Question
Appraisal System

Diagram:
- Investigator/Designer
- Survey Question
- Interviewer
- Respondent
- Background: Sociocultural/linguistic context

Relations:
- IAQ
- SAQ
Different methods target different subsets

Focus Group
Different methods target different subsets

Papers on Focus Groups:

- Cosenza: Focus group for existing survey: Thursday
- Levin: Innovative Approaches: Thursday
- Roller: Focus Group Workshop: Thursday
- Terry: Native Americans: Saturday
Different methods target different subsets

Use as many methods as possible (E. Geisen)

So, (I) Where are we?

- We have a wide variety of tools at our disposal, but not clear how to use them
- I’m not sure how much ‘scientific theory’ underlies this
Different methods target different subsets

Use as many methods as possible (E. Geisen)

II. Relevance: Do we still matter?
Does the world care about QDET?

- Sometimes seems that all the action is in non-prob samples, response rates, big data, etc...
- Do we care about the questionnaire?
- 6/6 CI respondents: Yes!
  - Clients, too – K. Levin: “Now more than ever” – especially once they see CI, are converts
  - J. Fowler: Butterfly Effect: Small changes in wording have huge downstream measurement effects
Does the world care about QDET?

• BUT: Are we at the cusp of a change?
• The world is full of well-meaning people who believe that anyone who can write plain English and has a modicum of common sense can produce a good...

Introducing:
A NEW AND IMPROVED 52" Buggy Whip
Key Question: Should we worry about how to ask questions?

If that train has run as far as it can... What else can we do?

1) Return to motivation as an explanation/leverage point – original Cannell, Krosnick (satisficing)
   o Increase attractiveness to Digital-Native-Multi-Taskers? →

---

**QUESTIONNAIRE**

- WHAT GENDER ARE YOU?
  - Male
  - Female

- HOW OLD ARE YOU?
  - 14–16
  - 17–19
  - 20–22
  - 23+

- HOW MUCH WOULD YOU PAY FOR A MUSIC MAGAZINE?
  - £2–£3
  - £4–£5
  - £6–£7

- HOW OFTEN WOULD YOU LIKE THE MAGAZINE TO BE RELEASED?
  - Monthly
  - Weekly

- WHAT WOULD YOU LIKE THE DOUBLE PAGE SPREAD TO BE ON?
  - An Interview
  - A review
  - A story

- WHICH TITLE DO YOU PREFER?
  - Intone
  - Recorded

- WHAT MUSIC GENRE DO YOU LIKE?
  - R&B
  - Pop
  - Indie
  - Hip-hop

- WOULD YOU LIKE THE MAGAZINE TO BE CONVENTIONAL OR UNCONVENTIONAL?
  - Conventional
  - Unconventional

- WOULD YOU LIKE TO SEE AN ESTABLISHED OR UNKNOWN ARTIST ON THE FRONT COVER?
  - Established
  - Unknown

- WOULD YOU LIKE THE MAGAZINE TO NAVIGATE YOU TO THE WEBSITE?
  - Yes
  - No

---

QDET2 - Willis Keynote
National Cancer Institute
What else can we do?

1) Make things more attractive, motivating...
2) Switch to devices that better reach respondents, and work for their brains... -> Callegaro Keynote, QDET2
3) Go to another data source: Q -> Non-Q
- Argument: We are moving from subjective to objective measures
  - No need for unreliable, unmotivated, hung-over survey respondents
  - Instead, get the real, objective data
  - Example: DNA changes (methylation alterations) could be used to reveal a person's smoking history, to better inform studies that explore risk factors for diseases like heart disease and lung cancer
Are we moving away from Q data?

– But, recent trend/revolution in biomedical/epi field:
  • No longer rely on doctor notes, records, superficial sources
  • Go right to the horse’s mouth, and get Patient Reported Outcomes (PROs), which are... self-report

Please respond to each item by marking one box per row.

(Questionnaire)

In the past 7 days...

<table>
<thead>
<tr>
<th>Question</th>
<th>Not at all</th>
<th>A little bit</th>
<th>Somewhat</th>
<th>Quite a bit</th>
<th>Very much</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAINING1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>PAINING2</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>PAINING3</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>
Are we moving away from Q data?
Records > Report? NOT a new debate

- Charlie Cannell -> Don Camburn (U MI)
  - Set up survey record-check study with docs
  - “This is great – we’ll finally know how good our records are.”

- Willson (NCHS): Cog Interviewing of birth certificates; Jobe, et al. (NCHS); CI of death certificates
  - From birth to death, there is error in the records

- Couper (2013): Self-report Q data are not going anywhere...
  - Challenge is when to use Q versus NON-Q
But: What about emerging non-Q data?

• Social media? Use of Facebook Posts/Tweets?
  – Not an either-or situation – social media may be useful in other ways than ‘data collection’

• *e.g.*, *ethnographic study in order to obtain initial input information*

• **Social Media = Big Data Focus Group?** *To learn about ‘contours and boundaries of the concept under study’*

• **How should QDET-ers look at Non-Q sources?**
Total Survey Error Components Linked to Steps in the Measurement and Representational Inference Process (Groves et al. 2004).

Robert M. Groves, and Lars Lyberg Public Opin Q
2010;74:849-879

© The Author 2011. Published by Oxford University Press on behalf of the American Association for Public Opinion Research. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com
Q versus Non-Q sources

<table>
<thead>
<tr>
<th>Measurement category</th>
<th>Q: survey self-report: ask questions</th>
<th>Non-Q: social media, device, records...</th>
</tr>
</thead>
</table>
| **Prob**             | - AAPOR standard model: garden variety probability survey  
| Probability sampling | - Probability-based Web panels        | Sampling based on demographics, etc.; processing of obtained (social media/record/device) data just as survey data |
| **Non-Prob**         | Non-probability surveys  
| Not sampled via probability mechanisms | - Web panels                           | Big Data sources  
|                      |                                      | - Mega Focus Group?  
|                      |                                      | - Pseudo-survey?    |

**QDET2 - Willis Keynote**
# Q versus Non-Q sources

## Measurement category

<table>
<thead>
<tr>
<th></th>
<th>Q: survey self-report: ask questions</th>
<th>Non-Q: social media, device, records...</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Prob</strong></td>
<td>- AAPOR standard model: garden variety probability survey - Probability-based Web panels</td>
<td>Sampling based on demographics, etc.; processing of obtained (social media/record/device) data just as survey data</td>
</tr>
<tr>
<td><strong>Non-Prob</strong></td>
<td>Non-probability surveys -Web panels</td>
<td>Big Data sources - Mega Focus Group? - Pseudo-survey?</td>
</tr>
</tbody>
</table>

## Representation category

- **Q**: self-report
- **Non-Q**: social media, device, records...

**Prob**: Probability sampling

**Non-Prob**: Not sampled via probability mechanisms
# Q versus Non-Q sources

## Measurement category

<table>
<thead>
<tr>
<th></th>
<th>Q: survey self-report: ask questions</th>
<th>Non-Q: social media, device, records...</th>
</tr>
</thead>
</table>
| **Prob**         | - AAPOR standard model: garden variety probability survey  
| Probability sampling | - Probability-based Web panels  | Sampling based on demographics, etc.; processing of obtained (social media/record/device) data just as survey data |

**Too narrow**

|                  | Non-probability surveys  
| **Non-Prob**     | - Web panels  | Big Data sources  
| Not sampled via probability mechanisms | | - Mega Focus Group?  
|                  | - Pseudo-survey? |
(II) Does QDET matter?

Of course…

But… but we may need to start thinking about ‘NonQ-DET’ as well
III. Where are we going?
How do we reach the Emerald City of Minimal Standards/Best Practices?
How do we tell a goodfrage from a bad frage?

*Fundamental question for QDET: D versus ET*

- Huge latent debate: (a) Rules/guidance versus (b) Empirical testing
- Debate subsumes several unresolved issues:
  - Value of *Expert Review* versus *Cognitive Testing*
  - *Absolute (static) versus Contextual (dynamic) notion of question quality*
• Presser, Couper, Lessler, E. Martin, J. Martin, Rothgeb, and Singer (2004): “…pretesting is the only way to evaluate in advance whether a questionnaire causes problems for interviewers or respondents.”

• This is the premise of Cognitive Interviewing
  – Collins (2015); Miller, Willson, and Padilla (2015); Willis (2005; 2015)
  – Kristen Miller (QDET2 cognitive interview):
    • Expert Review can’t tell you whether a question is double-barreled; we need to test!
    • Lots of researchers lack appropriate training in questionnaire evaluation
Short of agreeing to disagree... Do we even know that we disagree?

- But, how much is this view universally held?
  - Jen Dykema (QDET2 cognitive interview):
    (a) Academic center relies on expert review, design rules, literature, rather than cog interviewing
    (b) Practice is likely to be **sector specific**
    (c) Practices depend largely on **resources/mandates** rather than on science or theory
    (d) “Lots of people have no training **in questionnaire design**” (...as opposed to evaluation)

Dykema: Invited presentation on design based on question characteristics: Friday
Zavala-Rojas: Poster on Survey Quality Predictor (SQP): Thursday
Fundamental point of departure: Absolute versus Contextual notions of validity

• Not a new debate!

• **Design-Rule-Based = Absolutism:**
  – Meaning, and function, are inherent within the question
    • If it looks like a duck, it’s a duck! – So design the duck right
  – Leads to notions of “Database of good/bad questions”

• **Empirical = Contextualism:**
  – Oppenheim: Questions may have multiple validities, depending on objectives
  – Interpretive view: Meaning isn’t static, but is a function of the social interaction/expectation/context
  – Understanding function in context (usually) requires testing
Toilet Paper Model: Approach Depends on Objectives/Context

Why I insist on hanging toilet paper the "wrong" way
How do we resolve this debate?

• Beatty (QDET2 Cognitive Interview): “Both extremes are really really bad”
  – Rules alone will never predict outcomes
  – But, there is danger in mindless, unguided empiricism

• The appropriate balance is not clear
  – Bottom line: We should be arguing more about this!
  – We should NOT just go along on parallel tracks as governed by operational constraints and institutional mandates

• My hope is that QDET2 talks won’t just branch off into separate
  – Schools
  – Cliques
  – Echo chambers
  – Factions
So, (III) Where would I like to be going?

– Let’s better describe what we do (Boeije and Willis: CIRF)
– Let’s disseminate what we find (QBANK)
– Let’s communicate what we think works (Miller, Dykema, Fowler, Geisen…)
– Let’s be serious about criteria for declaring victory and going home!

  • Fowler: Cognitive Testing has been a clear victory
  • But: Don’t be TOO comfortable... Common
    – Cog testing is effective... because it improved
    – The questions are improved... because Cog testing improved
– To evaluate: We need an external measure of error/quality
  (so say all my cognitive interview subjects)
In closing...

1) Remember that: ‘The uncreative mind can spot wrong answers, but it takes a very creative mind to spot wrong questions.” - Anthony Jay

2) Enjoy QDET2

3) See you at QDET3 (2002… 2016… 2030?)