How do we Know Cognitive Interviewing is Any Good?

Gordon Willis, Ph.D.
Applied Research Program
Division of Cancer Control and Population Sciences
National Cancer Institute
Willisg@mail.nih.gov
How do we Know Cognitive Interviewing is Any Good?

Done by… who?

Gordon Willis, Ph.D.
Applied Research Program
Division of Cancer Control and Population Sciences
National Cancer Institute
Willisg@mail.nih.gov

Who is ‘we’?

Good for… what?
1) “We don’t need to prove anything, because cognitive interviewing is based on [fill]”

Where [ fill ] =

(a) Cognitive Theory (e.g. Tourangeau model)
(b) Qualitative Research Methodology

*I don’t think it’s that easy…*

2) So - we need to collect a body of empirical evidence to demonstrate method [reliability/validity/effectiveness]

Ok… how?
Developing a Framework for Evaluation

Wish I’d thought of that...


  (unlucky?) Chapter 13: Evaluation of Cognitive Interviewing Techniques -

  ➢ First, what evaluation question are we asking?

  Groves (1996): ‘How Do We Know What We Think They Think Is Really What They Think?’

  Nisbett and Wilson (1977): ‘Telling More Than We Know’

  *Are we really trying to be mind readers?*

  No! – We want to know *how survey questions function,* and we probe to get information relevant to that question

Gordon Willis NCI 3/2012
Models for the evaluation of cognitive interviewing (Willis, 2005)

A) Within-method evaluation:

Model 1) Demonstration of question improvement: Are questions improved by cognitive testing?

Model 2) Criterion validation: Are known problems found through cognitive testing?

Model 3) External validation: Are cognitive interviewing results replicated in the field environment?

Model 4) Reliability/Consistency analysis: Do independent cognitive tests, laboratories, or approaches identify the same problems?

Model 5) Process evaluation: Are cognitive interviewing results useful in the broad scheme of survey development?

B) Between-method evaluation:

Are the problems found in cognitive interviewing similar to those found by other pretesting methods?
Evaluation Model 1: Are questions improved by cog testing?

From Willis (2005): Linguistic Analysis of questionnaire, pre- and post-cognitive interviewing

<table>
<thead>
<tr>
<th></th>
<th>Long sentences</th>
<th>Big words</th>
<th>Average number words/sentence</th>
<th>Sentence complexity index (0-100)</th>
<th>Flesch-Kincaid reading level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial Draft</td>
<td>10</td>
<td>53</td>
<td>28.5</td>
<td>83</td>
<td>13.1</td>
</tr>
<tr>
<td>Recommended Draft</td>
<td>2</td>
<td>43</td>
<td>23.3</td>
<td>65</td>
<td>10.9</td>
</tr>
</tbody>
</table>

Looks good! End of story… (?)
Evaluation Model 1: Are questions improved by cog testing?

- Problem:

  Conrad & Blair (1996); Willis et al. (1999):
  If questions are improved, may be because designers are good at what they do. Who says we need cognitive interviewing?

  Willis (2005):
  Conversely, if questions are not improved, maybe the designers are (drunk / lazy / no good…)

  Difficult to separate (a) the process from (b) the staff incorporating it –

  “Cognitive testing doesn’t improve survey questions – questionnaire designers improve survey questions”
Evaluation Model 2: Criterion validation: Are known problems identified by cog testing?

- So, we focus on finding problems, rather than fixing them

Conrad & Blair have made some progress here: Embed ‘bad’ questions – do we find them?

- Challenges:
  - Difficult to identify ‘known bad questions’ from the point of view of a response error model
  - Assumes that ‘finding problems’ is our goal – what if we instead are interested in:
    (a) The tradeoffs associated with use of a particular question for a particular purpose (Beatty)
    (b) What a question ‘captures’ (Miller)
Evaluation Model 3: External validity: Do C.I. findings extend to ‘the field’?

Version 1: On a typical day, how much time do you spend doing strenuous physical activities such as lifting, pushing, or pulling?

Version 2: (a) On a typical day, do you spend any time doing strenuous physical activities such as lifting, pushing, or pulling?

(b) IF YES: ask Version 1

Prediction: For reports of 0, Version 1 < Version 2

<table>
<thead>
<tr>
<th>FIELD PRETEST (n=78)</th>
<th>0</th>
<th>&lt;1</th>
<th>1-4</th>
<th>5+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Version 1</td>
<td>32%</td>
<td>32%</td>
<td>35%</td>
<td>0%</td>
</tr>
<tr>
<td>Version 2</td>
<td>72%</td>
<td>18%</td>
<td>10%</td>
<td>0%</td>
</tr>
</tbody>
</table>

• Doesn’t ‘prove’ that the question is good/bad – but I like this approach
**Evaluation Model 4:**
Reliability: Do independent C.I. tests reveal similar results?

Five labs conducted interviews using own probing style, analysis procedures:

<table>
<thead>
<tr>
<th></th>
<th>English</th>
<th>Spanish</th>
<th>Chinese</th>
<th>Korean</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCI</td>
<td>16</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>25</td>
</tr>
<tr>
<td>Westat</td>
<td>18</td>
<td>36</td>
<td>9</td>
<td>9</td>
<td>72</td>
</tr>
<tr>
<td>NCHS</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>15</td>
</tr>
<tr>
<td>PHI</td>
<td>18</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>36</td>
</tr>
<tr>
<td>TOTAL</td>
<td>67</td>
<td>45</td>
<td>9</td>
<td>27</td>
<td>148</td>
</tr>
</tbody>
</table>
Please circle the single number (on a scale from 1 to 5) that best describes how concerned you feel right now about the following things:

**FEELINGS OF CONCERN NOW**

<table>
<thead>
<tr>
<th></th>
<th>Not at all</th>
<th>Hardly</th>
<th>Somewhat</th>
<th>Very much</th>
<th>Extremely</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Breast cancer occurring in me.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>2. My family’s history of cancer.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>3. What I can do to prevent breast cancer.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>4. Breast cancer hiding silently in my body.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>5. Not being able to avoid getting breast cancer.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>6. How I would feel if I had breast cancer.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>7. My chances of getting breast cancer in the future.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>8. Getting breast cancer without my knowing it.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>9. Finding out if I have breast cancer.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>10. What I can do to detect breast cancer early.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>11. Inheriting cancer from my “genes”.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>12. What having breast cancer would do to my body.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>13. Symptoms or signs of breast cancer in me.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>14. My chances of dying of breast cancer.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>15. How I would deal with breast cancer if I got it.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>16. Things I do that affect my risk of breast cancer.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>
How likely this is?  
How much this has occurred?  
Something else? (other than “How concerned I am”)  

Bottom line: Everybody found the same thing – results were very reliable
Summary: Is C.I. any good?

- This is amenable to empirical research
- There’s no single evaluation model that uniquely addresses the question
- QUEST members might consider how to collaborate in order to:
  
  (a) Develop evaluation models, criteria
  
  (b) Do fun, interesting, useful, publishable stuff