Current Developments in the Cognitive Testing of Survey Questions

Gordon Willis, PhD

Division of Cancer Control and Population Sciences
National Cancer Institute
National Institutes of Health
Rockville, Maryland
Willisg@mail.nih.gov

AAPOR Webinar: June 13, 2018

1

Outline for Webinar

My approach:

- Talk for 60-70 minutes
- 15 minutes for questions
- Suggest major developments, with
  (a) pointed examples
  (b) direction to further resources

2

Outline for Webinar

Major Developments I will describe:

A. Evolution of perspectives concerning 'what cognitive testing is'
B. Implications for sample size
C. Cognitive testing across survey administration modes and web platforms
D. Achieving cross-cultural comparability
E. Development of useful tools for cognitive testing

3

A. Perspectives on Cognitive Testing

- ***Useful sources:***
  - Willis (2005) – Standard, general guide
  - Miller, Willson, Padilla (2014) – Interpretive Perspective
  - Collins (2015) – Procedural guide
  - Willis (2015) – Analysis of the Cognitive Interview in Questionnaire Design

- *Note that a lot of action has been recent –*
  - *2005 – 2014: N = 1 book*
  - *2014-present N = 4 books*

4

Perspectives on Cognitive Testing: Reparative Versus Descriptive Testing

1) Reparative testing – Repair broken items!

- Prior to fielding, so pretesting orientation
- Emphasis on finding serious problems
- Tends to be quick, with small sample size
- Example:
  1) "Have you had your blood tested for the AIDS virus?"
  2) Vague: Did I actively do this, or was it done?
  3) "As far as you know, has your blood been tested for the AIDS virus"
- Focus is on ***reduction*** in response error

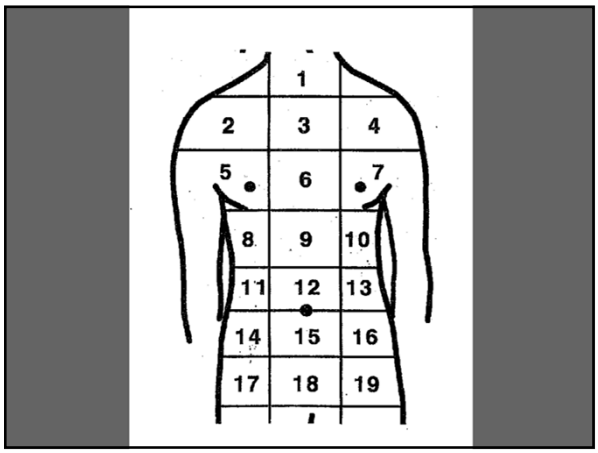
5

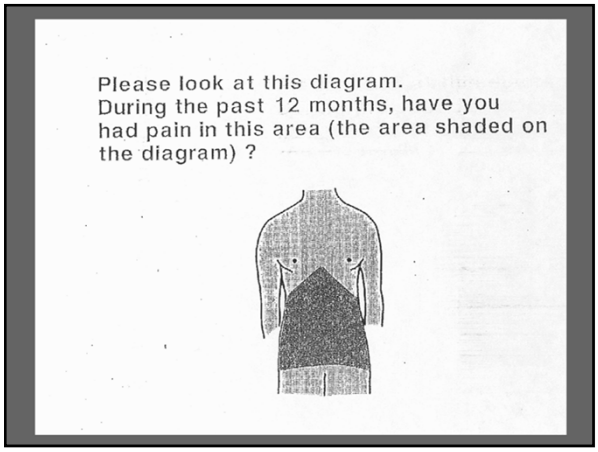
Tested ("classic") question: Pain in the abdomen

"In the last year have you been bothered by pain in the abdomen?"

What (Anticipated) probes make sense here?

- *What time period are you thinking about, exactly?*
- *What does "bothered by pain" mean to you?*
- *Where is your "abdomen?"*





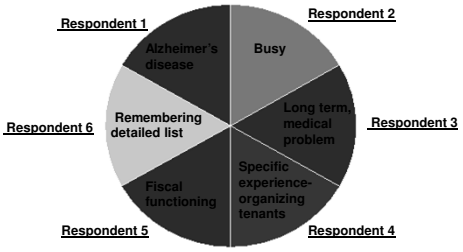
Perspective on Cognitive Testing: Reparative Versus Descriptive Testing

2) Descriptive testing – Understand question function

- Associated with *Interpretive* perspective (Miller, Willson, Padilla, 2014)
- May be either before, during, or after fielding
- Emphasis on understanding what the item measures – and what it doesn't
- Tends to use larger sample sizes -> Test to ***saturation***
- Example:
  1) "Would you say your health in general is excellent, very good, good, fair, or poor?"
  2) Probe: What does 'health in general' make you think of?
  3) Categorize results: Physical, Mental/Emotional/Spiritual…
  4) Assess variation in interpretation by subgroup (e.g. Hispanic)
- Focus is on ***measurement*** of response error

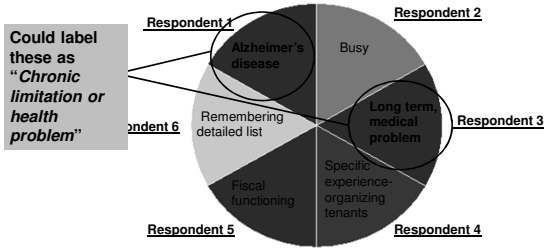
9

Overall, during the past 4 weeks, how much difficulty did you have with thinking clearly and solving daily problems?

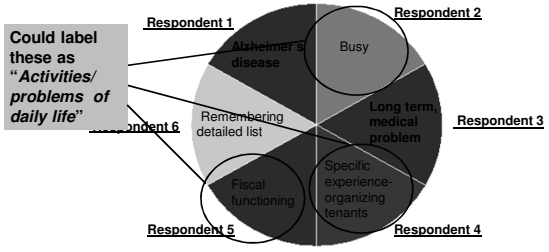


Source: Miller, K. (2008)

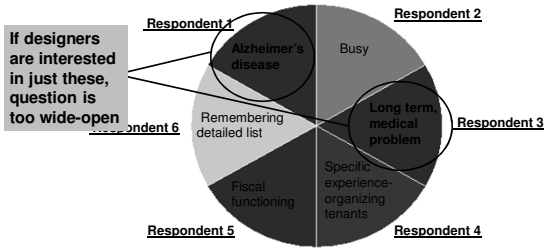
Overall, during the past 4 weeks, how much difficulty did you have with thinking clearly and solving daily problems?



Overall, during the past 4 weeks, how much difficulty did you have with thinking clearly and solving daily problems?



Overall, during the past 4 weeks, how much difficulty did you have with thinking clearly and solving daily problems?



If designers are interested in just these, question is too wide-open

B. Sample size: How many interviews do we need?

A major 'point of departure' in the field!

May depend what we are trying to accomplish:

- ***Reparative testing*:**
  - If we want to find ALL problems: Need a lot of interviews (50+) (Blair/Conrad)
  - If we want to improve questionnaire: We do what we can…
    May rely on VERY few interviews!

- ***Descriptive testing – what the question captures***
  - If we want a complete picture, we need to do enough testing to assess the full range of interpretations
  - NCHS (Paul Scanlon): In order to assess relative frequency of interpretation, we also need to select a representative (large) sample

14

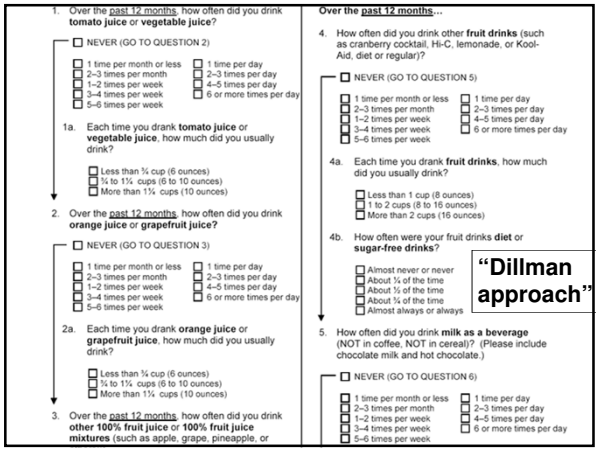
C. Adaptation to Mode and Administrative Platform

1) **Increased migration to self-administration (Dillman)**
2) **New technologies**

15

C1) Adapting cognitive interviewing to Self-administration (SAQ)

- The cognitive demands of SAQ are different from those of IAQ (face-to-face, phone:
  - SAQ relies on ***visual*** rather than ***auditory*** processing
  - SAQ sometimes requires ***navigational activities*** (find the starting point, follow skip patterns…)
- So, for the C.I., it may not make sense to follow the usual procedure of having the interviewer read the questions to the subject



1. Over the past 12 months, how often did you drink **tomato juice** or **vegetable juice**?

☐ NEVER (GO TO QUESTION 2)

☐ 1 time per month or less
☐ 2–3 times per month
☐ 1–2 times per week
☐ 3–4 times per week
☐ 5–6 times per week
☐ 1 time per day
☐ 2–3 times per day
☐ 4–5 times per day
☐ 6 or more times per day

1a. Each time you drank **tomato juice** or **vegetable juice**, how much did you usually drink?

☐ Less than ¾ cup (6 ounces)
☐ ¾ to 1¼ cups (6 to 10 ounces)
☐ More than 1¼ cups (10 ounces)

2. Over the past 12 months, how often did you drink **orange juice** or **grapefruit juice?**

☐ NEVER (GO TO QUESTION 3)

☐ 1 time per month or less
☐ 2–3 times per month
☐ 1–2 times per week
☐ 3–4 times per week
☐ 5–6 times per week
☐ 1 time per day
☐ 2–3 times per day
☐ 4–5 times per day
☐ 6 or more times per day

2a. Each time you drank **orange juice** or **grapefruit juice**, how much did you usually drink?

☐ Less than ¾ cup (6 ounces)
☐ ¾ to 1¼ cups (6 to 10 ounces)
☐ More than 1¼ cups (10 ounces)

3. Over the past 12 months, how often did you drink **other 100% fruit juice** or **100% fruit juice mixtures** (such as apple, grape, pineapple, or

Over the past 12 months…

4. How often did you drink other **fruit drinks** (such as cranberry cocktail, Hi-C, lemonade, or Kool-Aid, diet or regular)?

☐ NEVER (GO TO QUESTION 5)

☐ 1 time per month or less
☐ 2–3 times per month
☐ 1–2 times per week
☐ 3–4 times per week
☐ 5–6 times per week
☐ 1 time per day
☐ 2–3 times per day
☐ 4–5 times per day
☐ 6 or more times per day

4a. Each time you drank **fruit drinks**, how much did you usually drink?

☐ Less than 1 cup (8 ounces)
☐ 1 to 2 cups (8 to 16 ounces)
☐ More than 2 cups (16 ounces)

4b. How often were your fruit drinks **diet** or **sugar-free drinks**?

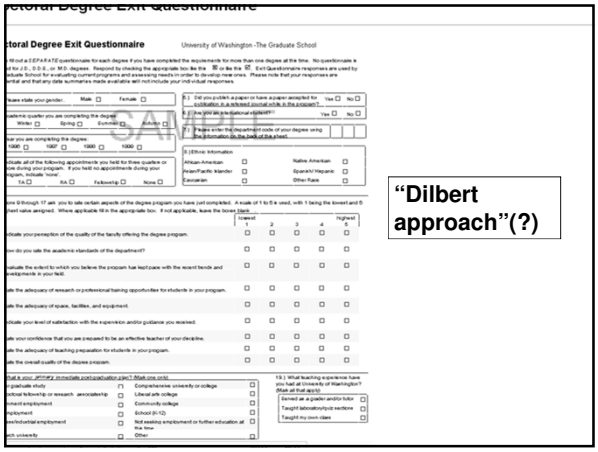
☐ Almost never or never
☐ About ¼ of the time
☐ About ½ of the time
☐ About ¾ of the time
☐ Almost always or always

"Dillman approach"

5. How often did you drink **milk as a beverage** (NOT in coffee, NOT in cereal)? (Please include chocolate milk and hot chocolate.)

☐ NEVER (GO TO QUESTION 6)

☐ 1 time per month or less
☐ 2–3 times per month
☐ 1–2 times per week
☐ 3–4 times per week
☐ 5–6 times per week
☐ 1 time per day
☐ 2–3 times per day
☐ 4–5 times per day
☐ 6 or more times per day



toral Degree Exit Questionnaire

toral Degree Exit Questionnaire University of Washington -The Graduate School

SAMPLE

"Dilbert approach"(?)

Self-administration: Concurrent versus Retrospective probing

- *There are several ways to conduct the probing:*
  - *Think-aloud, as the participant is completing the survey*
  - *Concurrent probing: 'Stop and probe'*
  - *Retrospective probing: Debriefing*
  - *Redline, et al.* (1998): 55-subject study comparing *think-aloud* with *retrospective probing*, for a paper questionnaire
  - Findings were very similar across approaches
    - Except, ***S's with low educational level tended to miss skip patterns under think-aloud***
    - Probably because they were focused on the verbal, so missed the visual/navigational
    - Authors recommended using ***both T/A and retro probing***, for SAQ questionnaires

If probing… Should probing be concurrent or retrospective?

- Has been a hiatus in conducting methods research on cognitive testing
  - Not much done since '80s, '90s
- More recently, advent of web panels with split samples has prompted re-emergence
- Fowler (2016) Cognitive Probe Placement –
  - Does it matter if probes are Concurrent versus Retrospective?

  Not really…

Marriage of Cognitive Testing and Usability …

- 'Tech-heavy' surveys need to be usable by user
- Recent convergence of CI, UT
- Example: Mary Davis, US Census Bureau:
  - Currently conducting 2 rounds of cognitive interviewing, then 2 rounds of joint cognitive and usability testing
  - Challenge: Balancing cognitive probes with probes on usability issues
  - "Fusion" of CI, UT should be done more, and earlier
  - *UT tends to be done after all CI has been completed and the questionnaire is "final" but necessitates further change to that version*

C2: New Tech--Video-conferenced cognitive interviewing (e.g., Skype)

Positive features:

- Includes, audio, video, and text capabilities
- Allows for enhanced geographic coverage
- Retains capacity for face-to-face interaction
- Plug-ins are available for recording (e.g., *Vodburner)*
- Is becoming commonplace for job interviewing, meetings

22

New Tech: Video-conferenced cognitive interviewing (e.g., Skype)

BUT:

- Requires a level of technological capacity and proficiency by the CI participant
  - YOU may need to provide IT expertise
  - You need to adjust for 'hiccups' in transmission
- Remove visual clutter behind interviewer
- Transmit image of interviewer's upper body, not just head
- Make sure interviewer doesn't blend in with background
- Emphasize eye contact by staring into camera
- Use a more animated voice than in face-to-face
- Practice first with a CI surrogate

C2: "Radical" C.I. Development: Web probing

- Embed cognitive probes in a web survey
- Concept is not new… Schuman (1966) "Random probe"; Converse and Presser: "Embedded probe"-

  Target item: *Government is trying to do too many things that should be left to individuals and private businesses*

  Probe: *"Can you tell me a little more about what you mean?"*
- More recently: Conduct pretesting using web probing– e.g., Amazon Mechanical Turk:
  - Instead of lots of info from a few people…
  - Get a little info from LOTS of people (> 1,000)

MSK-CC 5-2018

Example of Web-Based Probe
S. Fowler, et al,. 2015

We asked some questions about whether there are places in your neighborhood that you can walk to. Please say more about how you decided whether or not you can walk there?

Fowler, et al.: Most Responses to Probes Were Informative

"I would consider places **unwalkable** if there are **busy, high speed roads** and it would take **more than 10 minutes to walk** there."

"I deem a place within **walking distance** if I can walk there within **30 minutes**"

"**I live in a hilly area** so I was thinking about the fact that it's **uphill** to get to downtown (ha, up to get down), plus there's **no sidewalk and some blind corners**."

Fowler et al.- Example Web Probing Result

3	When answering the questions about places in your neighborhood that you can walk to, were you thinking of places that YOU might walk to, or places that OTHERS might walk to, or both?	Percentage of respondents providing responses that were about: Themselves: 40% Others: 2% Both: 37% Miscellaneous: 24%	Inter-rater reliability (Kappa) for judgement for the 4 categories: 0.886
4	When answering the questions about places in your neighborhood that you can walk to, were you thinking about what you actually DO, or about what you COULD do if you wanted to?	Percentage of respondents providing responses that were what they: Actually do: 48% Could do: 23% Both: 18% Miscellaneous: 11%	Inter-rater reliability (Kappa) for judgement for the 4 categories: 0.909

Fowler, et al. Perspectives on Web Probing: Advantages and Limitations

+ Get a BIG sample size, easily and cheaply
  (and, note that big sample size also allows for alternatives, like psychometric analysis)
+ Can conduct analysis according to demographic and other characteristics
+ Preliminary results suggest that responses are useful, and similar to those obtained one-on-one in the cognitive laboratory

BUT:

- Can select only a few items for probing
- The probes need to be fully scripted – inflexible
- There is no opportunity for follow-up probing

MSK-CC 5-2018

Scanlon and Edgar: Perspectives on Web Probing

	Traditional CI	Web Probing
Sample	Narrow Geographic & Demographic Dispersion	Wide Geographic & Demographic Dispersion
Goals	Explore and Evaluate	Explore and Evaluate or Confirm
Probes	Scripted or Reactive / Spontaneous	Scripted
Add'l Info	Personal Narratives Context	Response Distributions Paradata

Scanlon, P. and J. Edgar. 2017. "It's Not Either/Or: Web Probing and Cognitive Interviewing in Question Evaluation Studies." WAPOR Annual Meeting, Lisbon, Portugal. July 16, 2017

29

Close-Ended Web Probes (a la Paul Scanlon, NCHS):

- **Close-ended cognitive probes** can provide information about how respondents interpret survey items and arrive at their answers.
  - Require previous qualitative work
  - Less coding needed as compared to open-ended probes; easier for subgroup and other quantitative analyses
  - "Blend into" the survey—look like any other closed-ended survey item
- From the NCHS Research and Development Survey (RANDS) data so far, we find that:
  - The presence of close-ended probes does not appear to impact overall survey completion or item non-response rates
  - Close-ended probes do not appear to negatively impact the item-response rates of survey items that follow them
  - Respondents answer them at a slightly lower rate than non-probe items, particularly those formatted as "select all" questions with a large number of answer categories

National Center for Health Statistics
Center for Questionnaire Design and Evaluation Research

Scanlon, P. "Methodological Considerations for the Use of Close-Ended Web Probes." 7th Conference of the European Survey Research Association, Lisbon, Portugal. July 20th, 2017.

D. Achieving cross-cultural comparability

- **Language translation and cultural adaptation are increasingly important**
- **Back-Translation is no longer viewed as a gold standard for translation:**
  1) **"Are you feeling blue?" ->**  **?**
  2) **"Azul" ->**
  3) **"Blue"**  **?**
- **MUCH more reliance on empirical study of cross-cultural comparability**

31

Cross-Cultural Cognitive Interviewing (CCCI)

The application of cognitive interviewing to evaluate questionnaires for multiple *languages* and *cultures* would seem to be a natural extension (Chan and Pan 2011; Willis and Miller 2011)

But – there has been some question about how effective CCCI is in multilingual and multicultural contexts

I decided to check the published sources for evidence, once way or another:

Willis, G.B (2015), The Practice of Cross-Cultural Cognitive Interviewing. Public Opinion Quarterly. 79 (S1): 359-395. *doi: 10.1093/poq/nfu092*

32

Conclusion (1): CCCI has been applied widely

- **Languages:** English, Spanish, Chinese (Mandarin, Cantonese), Korean, Vietnamese, Thai, Bangla, Malaysian, Danish, French, Russian, German, Dutch, Hungarian, Bulgarian, Portuguese, and Maori;
- **Countries:** Mainly in North America but also Europe, Asia, Africa, Mexico, New Zealand, India, Bangladesh, Central America, South America
- **Constitutes evidence of "Diffusion of Innovation"**

33

Conclusion (2): CCCI studies have 'large' samples

- Due to the complexities introduced by increasing number of subgroups to be compared, CCCI studies often include numbers well beyond the 'textbook' range of 15–30 (e.g., Willis, 2005)
- Of the 31 listing total sample size:
  - **6 (19.4 percent) contained 30 or fewer participants**
  - **12 (38.7 percent) contained 31–100**
  - **13 (44.8 percent) contained 101+**

34

Conclusion (3): CCCI should include source language testing

- Translation assessment requires source-language cognitive interviews (*Carter, Schoua-Glusberg, and Sha 2009; Goerman and Caspar 2010*)
- Because: Source language version may contain problems: *Problems in the source language Fitzgerald et al. 2011*; C*ross-cutting* problems (*Goerman and Caspar 2010b*); G*eneric* problems (*Levin et al. 2009*)
- So, if we don't include CI in source interview, there is no baseline by which to evaluate the results related to the translation

35

Conclusion (4): CCCI findings are often not attributable to language or culture

- Group membership (language, culture) is often confounded with ***demographic variables***
- A few CCCI studies have attempted to control demographic factors: Saleska, Kanya-Ngambi, and Alvarado, 2009; Berrigan et al., 2010, Miller et al., 2005
- e.g.: Miller, et al. (2005): Age, rather than Hispanicity, accounted for the dominant effects noted
- However, the vast majority of qualitative CCCI studies have lacked such controls
- **Therefore, cross-cultural differences observed may be driven by confounding (age, educational level…)**

36

Conclusion (5): The interviewer matters

- Demographic/Personality/Experiential characteristics of interviewers drive the results
- We DO know that 'being bilingual' is not enough!
  - Attempted work-around of reading from script tends to fail
  - Flexible interviewing tends to work best, unless a lot of prior work is done to 'test the (standardized) probes'

37

Conclusion (6): Probes matter

- Has been suggested that non-Western/English speakers do not respond well to the cognitive interview:
  - Thinking aloud difficult; Probes are awkward/artificial/resisted
- Systematic review points to a different picture:
  - Think-aloud is generally difficult for those with low educational levels, and for those who lack 'survey literacy'
  - Probes that fail in CCCI studies are the same ones that do not work well in standard CI!
    - **Paraphrasing; Hypothetical probes**
  - Probes that function well are the same ones that do well in standard CI
    - **Meaning probes; Elaborative probes ("Tell me more")**

38

Conclusion (6): Probes matter – for ALL C.I.

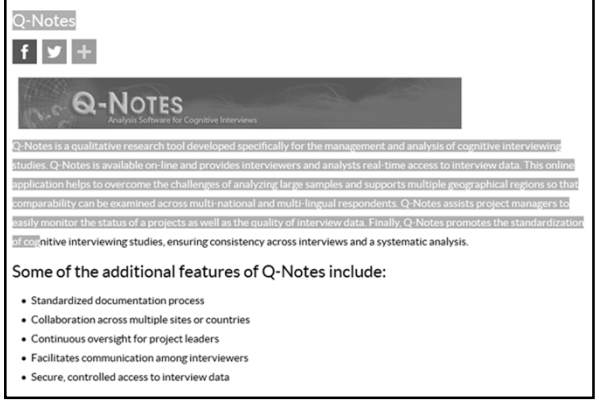
Fred Conrad, U MI: 2018 AAPOR presentation:

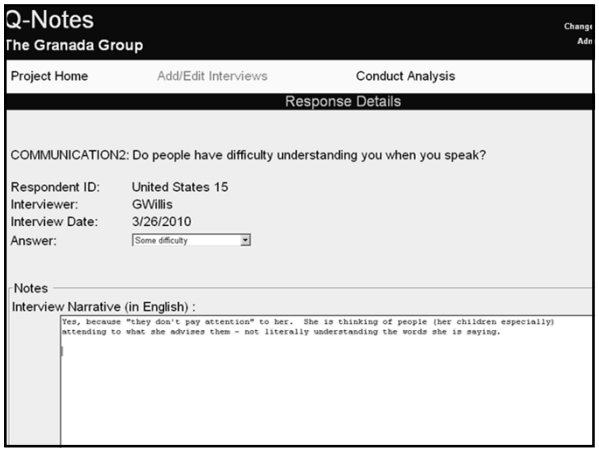
- Asking Cognitive Interview *R*s about *specific problems* or interpretations leads to (biased) affirmation of (non-existent) problems
- We should probe in *open-ended way*, even if there is reason to believe *R*s have experienced a particular problem
  - *May produce less "evidence" of problems but the evidence that is produced is likely to be more credible*
- Communicating to *R* the value researchers place on candid – not polite – responses may help…
- ***Empirical evidence for view that it's best to probe generally, i.e., "Tell me more…"***

39

E. New Tools for Cognitive Testing: Online and checklist systems

- **Q-Notes: Online analysis software specially designed for cognitive interviews:**
  http://www.cdc.gov/qdrl/b4product/prod220.htm
- **Cognitive Interview Report Format (CIRF)**
  Boeije & Willis (2013)
- **Q-Bank database of cognitive testing results**
  http://wwwn.cdc.gov/qbank/Home.aspx





Yes, because "they don't pay attention" to her. She is thinking of people (her children especially) attending to what she advises them - not literally understanding the words she is saying.

Conclusion: "Understanding you" is interpreted not as speech understanding, but interpreting a message/direction and then complying with it

Writing up cognitive testing results

- **There is no standard format**
- **Cognitive testing publications have lots of holes!**
  - **Boeije & Willis (2013):** ***Introduce Cognitive Interviewing Reporting Format*** **(CIRF)…**
  - ***In Special Issue of (journal) Methodology***
- **Ten major elements, with listing of info to be included under each:**
  - Some elements are normally omitted from C.I. reports
  - Some elements are included but under-specified

Cognitive Interviewing Reporting Format (CIRF)

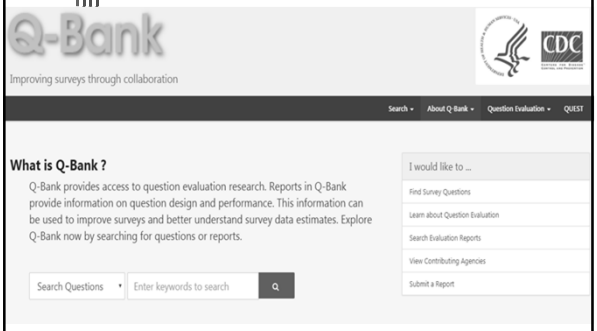
CIRF Element
(1) [ ] Research Objectives
(2) [ ] Research Design
(3) [ ] Ethics
(4) [ ] Participant Selection
(5) [ ] Data Collection
(6) [ ] Data Analysis
(7) [ ] Findings
(8) [ ] Conclusions, Implications, and Discussion
(9) [ ] Strengths and Limitations of testing process
(10) [ ] Report Format →

Boeije, H., & Willis, G. (2013). *The Cognitive Interviewing Reporting Framework (CIRF):* Towards the Harmonization of Cognitive Testing Reports. Methodology, 9(3): 87-95.

Q-Bank: Accessing Cognitive Test Findings

- Historically, cognitive test findings have been relatively inaccessible.
- Implications
  - Knowledge is lost
  - Resources are wasted
  - Lack of transparency and accountability
- Important need for a medium to make findings available

https://wwwn.cdc.gov/QBANK/Home.aspx



Q-Bank example: Search for 'smoking'

Survey	Question Text	Test Date
Not Applicable	During the past month, did you smoke any cigarettes, cigars, or a pipe?	2004
Not Applicable	Compared with the amount you usually smoke, was the amount you smoked in the past month more than usual, about the same, or less than usual?	2004
Not Applicable	Was the amount a lot, some, or only a little (more/less) than usual?	2004
NHIS	What was it that burned/scalded {you/participant name}? If response is fire or smoke ask: What caused the fire/smoke?	2002
NHIS	Have you smoked at least 100 cigarettes in your ENTIRE LIFE?	2003
NHIS	How old were you when you FIRST started to smoke fairly regularly?	2003
NHIS	Do you NOW smoke cigarettes every day, some days or not at all?	2003
NHIS	How long has it been since you quit smoking cigarettes?	2003
NHIS	On the average, how many cigarettes do you now smoke a day?	2003
NHIS	On how many of the PAST 30 DAYS did you smoke a cigarette?	2003
NHIS	On the average, when you smoked during the PAST 30 DAYS, about how many cigarettes did you smoke a day?	2003
NHIS	DURING THE PAST 12 MONTHS, have you stopped smoking for more than one day BECAUSE YOU WERE TRYING TO QUIT SMOKING?	2003

Select report for “How many times have you tried to quit smoking” ->

5. How many times have you tried to quit smoking?

_____

Two themes emerged from the evaluation of this question. First, respondents employed a number of calculation strategies when answering this question, including counting, estimating, and providing “random” numbers. Second, respondents also varied in terms of what they considered to be a quit attempt. This could range from anywhere between simply thinking about quitting smoking (without actually attempting to quit) to making a concerted effort to quit smoking. These findings are discussed below.

In closing…

‘The uncreative mind can spot wrong answers,

but it takes a very creative mind to spot wrong questions.”

- Anthony Jay