THINKING ALOUD

USE OF A RESEARCH TECHNIQUE WITH PHARMACY STUDENTS AND QUALIFIED PHARMACISTS

Hannah Family, Dr Jane Sutton & Prof Marjorie Weiss
Department of Pharmacy and Pharmacology
HSRPP Conference 2012 ~ 24/4/12 ~ Cork
OUTLINE

1. Questionnaire design and measurement error
2. Cognitive Psychology and questionnaire design
3. Introduction to the Think Aloud method
4. Our Think Aloud study
5. Our findings
6. How we resolved the measurement errors we found
7. Conclusions

The Think Aloud method helps you to identify measurement error and more importantly it can explain why it’s happening.
THE ASSUMPTIONS WE ALL MAKE WHEN WE USE QUESTIONNAIRES:

1. All respondents understand the questions in a consistent way

2. The questions are asking for information respondents have and can retrieve

3. The wording of the questions provides respondents with all the necessary information they require to be able to answer them in the way required by the researcher


But if these assumptions are not fulfilled your results will be wrong!
THE COGNITIVE PROCESSES INVOLVED IN ANSWERING A QUESTION

Answering questions is an inherently cognitive process

Collins (2003)
THE THINK ALOUD METHOD

• Also known as protocol analysis

• The development of this method is largely credited to Simon, Newell & Ericsson (see Newell & Simon, 1972, Ericsson & Simon, 1980)

• In the 1990s it was first proposed as a method for piloting questionnaires (Bolton, 1991)

• There are two forms of think aloud, concurrent and retrospective
THE PILOT STUDY (for)
THE EFFECTS OF MENTAL WORKLOAD ON COMMUNITY PHARMACISTS’ ABILITY TO DETECT DISPENSING ERRORS

Pilot study conducted Oct ‘10 – Feb ’11

Aim of the think aloud study: The aim of this study was to use the think aloud technique to assess the reliability and validity of two questionnaires used as part of a study into pharmacists’ workload.

Method:
• Participants recorded their own think alouds using a digital dictaphone
• They were given a practice session to get used to “thinking aloud”
• Think alouds were transcribed verbatim and analysed using QSR Nvivo 9

Sample for think aloud study
• 19 participants in total (16 pharmacy students & 3 community pharmacists)
• A further 6 community pharmacists took part in the re-pilot of the questionnaires
THE QUESTIONNAIRES

THE BIG FIVE PERSONALITY INVENTORY

- Measure of personality
- 44 items
- 5 point Likert scale
- Available in 10 different languages

Reliability / Validity data to date:
- Item analysis
- Factor analysis of items
- Convergent validity (comparison of 3 big five personality measures)
- External validity (via peer ratings)

Website: http://www.ocf.berkeley.edu/~johnlab/bfi.htm

THE DUNDEE STRESS STATE QUESTIONNAIRE

- A battery of questionnaires which assess the 11 primary dimensions of mood, motivation, and cognition in performance settings
- 96 items
- 4, 5 & 11 point Likert scales

Reliability / Validity data to date:
- Item analysis
- Factor analysis of items
- External validity (experimental vs field study data)

(Ref: Matthews et al 1999, 2002)

Both questionnaires are extensively validated
THINK ALOUD ANALYSIS & ASSESSMENT CRITERIA

Coding Schemes
1. Coding comprehension issues (Cannell, Fowler & Marquis, 1968):
2. Coding for retrieval issues (Bolton, 1991)
3. Coding for judgement issues (Sudman et al, 1996; Collins, 2003)
4. Coding for response issues (Sudman et al, 1996)

Assessment Criteria
1. Oskenberg et al (1991) criteria – if 15% or more of the sample experience difficulties with an item it needs reviewing

2. As we had 2 slightly different samples, pharmacy students and community pharmacists in practice, we required evidence of problems with an item in both samples before we reviewed the item.
FINDINGS:
BIG FIVE PERSONALITY INVENTORY

“Is ingenious a deep thinker well I didn’t think ingenious meant deep thinker, I didn’t think they were the same thing [chuckles] what happens if I put neither I don’t really know what it means.”

Pharmacy student, Participant 8. BFI item 15.
FINDINGS:
DUNDEE STRESS STATE QUESTIONNAIRE

4 out of the 96 DSSQ items were found to cause measurement errors for ≥15% of the sample.

1. Mood State Questionnaire - Item 7
   “At the moment I feel passive”
   
   37% of the participants did not understand the word “passive” in this item.

2. Mood State Questionnaire - Item 19
   “At the moment I feel unenterprising”
   
   53% of the participants did not understand the word “unenterprising” in this item.
FINDINGS:
DUNDEE STRESS STATE QUESTIONNAIRE

3. Motivation Questionnaire - Item 12
“I feel apathetic about my performance”

26% of the participants did not understand the word “apathetic” in this item.

4. Thinking style questionnaire – item 17
“I feel I have less scholastic ability than others”

21% of the participants did not understand the word “scholastic” in this item.

“I feel apathetic - apathetic I’m quite embarrassed now because I don’t actually know what that means so I’m going to leave it out”
Qualified pharmacist participant 22

“I feel that I have less scholastic ability right now than others [pause] don’t know what that means so I’ll put a question mark next to that because I don’t know what scholastic ability means”
Pharmacy Student Participant 11
FINDINGS: COMBINING THE BFI & DSSQ

We found some unexpected results for three of the sub-scales on the DSSQ

<table>
<thead>
<tr>
<th>Scale</th>
<th>Subscale</th>
<th>Time 1</th>
<th>Time 2</th>
<th>Time 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thinking style</td>
<td>Self-focused attention</td>
<td>10.60</td>
<td>4.60</td>
<td>2.30**</td>
</tr>
<tr>
<td></td>
<td>Self-esteem</td>
<td>18.40</td>
<td>22.20</td>
<td>22.50*</td>
</tr>
<tr>
<td>Thinking content</td>
<td>Task-irrelevant thoughts</td>
<td>15.10</td>
<td>10.50**</td>
<td>11.65*</td>
</tr>
</tbody>
</table>

* Significant at p<.05, ** significant at p<.001

Results of the DSSQ change over time Why...?
FINDINGS:

“I thought about something that happened in the distant past I guess in answering the previous questions I did so umm often”

Pharmacy Student, Participant 18, Thinking content item 15

“Umm I thought about my level of ability umm only when prompted to by the questionnaire so a few times”

Pharmacy Student, Participant 17, Thinking content item 5

Even validated questionnaires can show measurement problems if the cognitive context is not appreciated
WHAT WE DID NEXT

1. Contacted the author of the DSSQ – Prof Gerry Matthews – for his advice on how to reduce the measurement error for these items

- Prof Matthews suggested adding definitions as footnotes for the items that were causing comprehension issues.
  - This avoided the need for changes to wording of a pre-validated questionnaire (this means we can still compare our results to studies that have used this questionnaire)
  - This avoided changing the layout & appearance of the scale
  - He provided us with definitions that explained the intended meaning of the word

2. Re-piloted the questionnaire with a further 6 participants to check that the definitions rectified the measurement error

- No one in the re-pilot experienced the same difficulties

3. Changed the study procedure so the BFI was completed the day before the experiments and the DSSQ
CONCLUSIONS

1. Even validated questionnaires can show measurement problems – pilots and think aloud studies should be carried out for existing as well as new questionnaires.

2. Researchers and respondents need a shared understanding of the meanings inherent in questionnaire items.

3. When using two separate questionnaires in conjunction, especially if they are pre-existing, validated questionnaires, check for reactivity effects.

The Think Aloud method helps you to identify measurement error and can explain why it’s happening
A CAVEAT

The think aloud method is a data collection tool too. It will also be susceptible to certain types of measurement error:

“..feel self-conscious a little bit because I am talking out loud err [pause]”
Pharmacy student,
Participant 1, Thinking style item 11

Even the Think Aloud method can alter the cognitive context and impact the respondents’ answers
ACKNOWLEDGEMENTS & THANKS TO...

- **My Supervisors:** Dr Jane Sutton & Prof Marjorie Weiss
- **The rest of the pharmacy practice team at the University of Bath, with particular thanks to:** Mr Chris Coy & Dr Lynette James
- **The Pharmaceutical Trust for Educational and Charitable Objects**
- **Our project management team**
- **The participants**

For these slides and more information about this our research into the relationship between mental workload and dispensing errors visit:

[http://errorgirl.com](http://errorgirl.com)
REFERENCES


EXTRA SLIDES
THE THINK ALOUD METHOD

- Also known as protocol analysis
- The development of this method is largely credited to Simon, Newell & Ericsson (see Newell & Simon, 1972, Ericsson & Simon, 1980)
- In the 1990s it was first proposed as a method for piloting questionnaires (Bolton, 1991)
- There are two forms of think aloud, concurrent and retrospective

Table 1: the types of data the two think aloud methods provide (Sudman et al, 1996)

<table>
<thead>
<tr>
<th>Method</th>
<th>Cognitive Process</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Comprehension</td>
</tr>
<tr>
<td>Think Aloud</td>
<td></td>
</tr>
<tr>
<td>Concurrent</td>
<td>x</td>
</tr>
<tr>
<td>Retrospective</td>
<td>x</td>
</tr>
</tbody>
</table>

Diagram:
- Comprehension
  - Response
  - Retrieval
  - Judgement

Table:
- Method: Concurrent, Retrospective
- Cognitive Process: Comprehension, Retrieval, Judgement, Response
- x indicates presence of data
- X indicates absence of data

Diagram:
- Comprehension
  - Response
  - Retrieval
  - Judgement

Diagram:
- Comprehension
  - Response
  - Retrieval
  - Judgement
THE THINK ALOUD METHOD

**CONCURRENT**

Participant thinks out loud as they are solving a problem / answering a question

**Pros:**
- No interference from researcher
- Less time consuming overall
- More realistic if your questionnaire isn’t interview administered

**Cons:**
- Less detailed information
- May miss some problems

**RETROSPECTIVE**

Participant solves a problem/ is asked a question by an interviewer and once they give their answer **cognitive probes** are used to find out how they got to this answer

**Pros:**
- More detailed data than concurrent think aloud
- Cognitive probes will encourage quieter respondents to talk

**Cons:**
- Interviewer probes could cause contamination of results

“Some respondents, usually those with higher levels of education and greater verbal facility find the concurrent think aloud an easy and interesting task, others however need prompting turning a concurrent think aloud into a retrospective one.” Sudman et al 1996 p 34.
COMPREHENSION

The first thing we need to find out is simply whether the respondent has interpreted the questionnaire item in the way it was intended. This involves reading the transcripts to identify where:

1. the respondent requests clarification or says things like “I don’t know what that means (code as respondent needs more information)

2. The respondent has misunderstood the question without realising

This process was adapted from Cannell, Fowler & Marquis (1968)

I also identified one other code in the transcripts from my pilot, this was the use of conflicting terms, there were several instances of participants reporting that two words in the same question meant conflicting things to them.

Oskenberg, Cannell and Kalton (1991) suggest an arbitrary index of 15% to identify problem items. That is if 15% of your respondents have a problem with an item then this suggests this item is highly problematic.
“Is outgoing sociable umm [pause] umm yeah I’m kind of sociable not really outgoing again in the middle”

Participant 17
BFI item 36

“sometimes shy and inhibited [pause] inhibited [pause] umm I’d say I’m shy but not inhibited [pause] [sniff] ohh mmm yeah”

Participant 14
BFI item 31
Having comprehended the question, the respondent now needs to retrieve the relevant information from their long term memory.

Several factors can affect the retrieval of information

Factual Information
1. Whether or not the retrieval and encoding context match
2. How rare or distinctive the event was
3. Previously cued information (from questions, or events going on at the moment)

Attitudinal information

Here Sudman et al (1996) and Collins (2003) depart as Sudman argues that attitudes are not per se part of long term memory that they are based on judgements of events and therefore attitudes themselves are not retrieved.

“Is sometimes shy, inhibited yeah like I said earlier it depends on whose around me if I don’t know them then I suppose I’m quite shy and quiet and I don’t really let my personality come out so I will agree strongly with that one [pause]”

Participant 3, BFI item 31

“My thought about something that happened in the distant past I guess in answering the previous questions I did so umm often”

Participant 18, Thinking content item 15

“Depressed well I just said I was happy so I can’t be depressed as well can I”

Participant 3

UWIST item 14
RETRIEVAL: HUMANS ARE COGNITIVE MISERS

Unfortunately for researchers who conduct surveys, as well as all the factors mentioned before humans are cognitive misers and will usually follow a satisficing rather than an optimising information retrieval strategy.

This means that we will not search for every stored instance of an event we will search for a few and form a judgement based on that. If people cannot find one instance of an event quickly, they are unlikely (in the context of answering a questionnaire) spend ages racking their brains to find the event, or answer they need to answer your question.

This has big implications for the robustness of questionnaire data

“Has a forgiving nature mm I suppose that’s about other people’s opinion I’m not really sure about that myself mmm forgiving nature oh I’ll just say three”

Participant 17, BFI Item 17
RETRIEVAL: CODING OF THINK ALOUDS

How best to code concurrent think alouds for retrieval issues is not particularly clear. Sudman et al (1996) suggest that it is impossible to create a universal coding scheme because the topics of the questionnaire will vary and have an impact on the types of issues you are looking for.

For example, to be particularly awkward the questionnaire I have used asks lots of questions asking participants to think about the last 10 minutes. This creates problems as this information is likely to be in short term memory or lost (i.e. so inconsequential that it was not stored in long term memory).
BOLTON’S CODING SCHEME (1991)

“I don’t know about my enthusiasm and generating it for other people [pause]”  
Participant 1, BFI item 16

“Has an active imagination so mmm I don’t really know with that one [pause] I think because I don’t really know I’ll choose neither agree nor disagree there”  
Participant 5, BFI item 20

EXHIBIT TWO
Coding Categories

1) Comprehension - Similar
    identical&question
    same&question
    similar&question
    identical&answer
    same&answer
    similar&answer
    answer&again
    sound&alike
    question&close
    sound&identical
    I answered&that&question
    like:I said
    like:I said&before
    experience&again
    opinion&again
    there&again
    is that what you’re asking
    repetitious

2) Retrieval - No Experience
    not:experience
    never:experience
    not:experienced
    any:experience
    never:experience
    not:experience
    haven’t:experience
    not:familiar:with
    none:for:service
    none:for:option
    never:use
    never:done
    never:used
    don’t:use
    haven’t:used

3) Retrieval - Don’t Remember
    dont:remember
    forget
    can’t:think
    I’m:trying:to:think
    of: something

4) Judgment - Can’t Judge
    I can’t say
    I can’t tell
    I can’t rate
    I can’t:evaluate
    I can’t:judge
    tough&rate
    not:easy&rate
    difficult&rate
    hard&rate
    tough&evaluate
    not:easy&evaluate
    difficult&evaluate
    hard&evaluate
    tough:to:say
    not:easy:to:say
    difficult:to:say
    hard:to:say
    tough:judge
    not:easy:judge
    difficult:judge
    hard:judge
    tough&to:tell
    not:easy&to:tell
    difficult&to:tell
    hard&to:tell

5) Judgment - Uncertain
    approximately
    perhaps
    kind:of
    unless
    somewhere:in:there
    I reckon
    not:certain
    I imagine
    depends
    mostly
    sort:of
    not: sure
    whatever

* The string "word1:word2" is interpreted "word1 immediately followed by word2 in a given segment."  
The string "word1&word2" is interpreted "word1 and word2 occurring in a given segment," where word 1 and word 2 can appear in any order, with or without intervening words.
RETRIEVAL RESPONSE LATENCY MEASURES

Bolton (1991) also used a response latency measure with concurrent think aloud data. All pauses longer than 3 seconds for an item were taken as an indicator of retrieval difficulties.

Broken utterances (Sudman et al, 1996) were also used as indicators of retrieval difficulties.

I feel smart as others [long pause] it depends who I am comparing myself to Participant 1, Thinking style item 12.

“Energetic err umm yeah I’d give myself a three for that I think oh no a two maybe slightly I’m not sure actually yeah a two” Participant 3 UWIST item 3
JUDGEMENT: CONTEXT EFFECTS

- Response Alternatives
  - Open / closed question format (e.g. Sudman et al., 1996)
  - Scale range issues
    - Frequency scales (Schwarz, Strack, Müller & Chassein, 1988)
    - Likert-type scales (Schwarz, Knäuper, Hippler, Noelle-Neumann & Clark, 1991)

- Question context
  - Ambiguity (Strack, Schwarz & Wänke, 1991)
  - The “given new” (Schwarz, Strack, & Mai, 1991)
“can be moody [pause] [sigh] moody’s a difficulty one because sometimes I’m grumpy [chuckles] is that the same thing [pause] umm”
Participant 8, BFI item 29

“I thought about the difficulty of the problems mmm I suppose I thought the checking part coming up but [pause] does that count as a difficulty I suppose it does once”
Participant 8, Thinking content item 4
“Artistic interests mmm agree a few umm so artistic interests is that musical so I’d agree a little I wouldn’t say loads maybe one or two [pause] a few I don’t know what few would be there one or two two or three if you’d said three maybe you’d disagree but as a few I’d agree a little as I have one or two yeah”
Participant 5, BFI item 41

“Hmm [pause] am I someone who’s talkative umm [pause] I think I can be [pause] in the right environment yeah I think I am talkative to people that I know [pause] less so if it’s a stranger probably [pause] so I would say I agree I wouldn’t say agree strongly I’m not very very talkative but I’m not quiet so I’d say four but seems quite high”
Participant 7, BFI item 1
JUDGEMENT: WHY DO CONTEXT EFFECTS OCCUR?

It is rare that questionnaire respondents have the answer to your question pre-formed in their memory.

This means that most answers to survey questions reflect judgements made by respondents have generated on the spot in the context of the questionnaire.
JUDGEMENTS: JUDGEMENTAL HEURISTICS

“cognitive shortcuts”

A number of different strategies for estimating answers to frequency questions are known:

1. Recall of specific events
2. Estimation based on recall of summary information about the rate of occurrence of the event
3. Recall of an exact count of events
4. Estimation based on a general impression


Availability Heuristic (Tversky and Kahneman, 1973) – people estimate the frequency, likelihood or typicality of events by the ease with which they can bring relevant examples to mind.
## SUDMAN ET AL’S (1996) JUDGEMENT CODING SCHEME

<table>
<thead>
<tr>
<th>Automatic response</th>
<th>Counting strategies</th>
<th>Rate-based estimates</th>
<th>Enumeration-based estimates</th>
<th>Anchoring strategies</th>
<th>Miscellaneous for attitude questions</th>
<th>Search strategies</th>
<th>Event cues</th>
<th>Reference period</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automatic</td>
<td>General recall and count</td>
<td>General rate based estimation</td>
<td>General enumeration-based estimation</td>
<td>Same as self</td>
<td>Based on specific behaviour/event</td>
<td>No order/search</td>
<td>Person mentioned</td>
<td>Anchor date on public event</td>
</tr>
<tr>
<td>Event did not occur/non-event</td>
<td>Counting with adjustment for uncertainty</td>
<td>Rate based estimation with adjustment based on specific incident (addition/s to estimate)</td>
<td>Enumeration-based estimation with adjustment based on specific incident (addition/s to estimate)</td>
<td>Based on prior answer</td>
<td>Based on discussions with other</td>
<td>Forward search</td>
<td>Place mentioned</td>
<td>Anchor date on personal event</td>
</tr>
<tr>
<td>Retrieval of prior judgement</td>
<td>Counting with expression of uncertainty (no adjustment)</td>
<td>Rate-based estimation with adjustment based on general knowledge</td>
<td>Enumeration-based estimation with adjustment based on general knowledge</td>
<td>Anchor on self and adjust</td>
<td>Based on general knowledge</td>
<td>Backward search</td>
<td>Emotional reaction to event mentioned</td>
<td>Anchor date on season of the year</td>
</tr>
<tr>
<td>No probe therefore no protocol</td>
<td>Counting by domain</td>
<td>Rate-based estimation with adjustment based on non-occurrence (subtraction/s from estimate)</td>
<td>Enumeration-based estimation with adjustment based on non-occurrence (subtraction/s from estimate)</td>
<td>Anchor on norm and adjust</td>
<td>Based on attitude towards issue</td>
<td>Anchor on date and forward search</td>
<td>Time of event occurrence mentioned</td>
<td>General characteristic of event/person</td>
</tr>
<tr>
<td>Don’t know</td>
<td>Counting by domain with adjustment for uncertainty</td>
<td>Rate-based estimation with adjustment for uncertainty</td>
<td>Enumeration-based estimation with adjustment for uncertainty</td>
<td>Anchor on another specific person and adjust</td>
<td>Anchor on date and backward search</td>
<td>Based on prior response (one answer triggers another)</td>
<td>“Always…”</td>
<td></td>
</tr>
<tr>
<td>Not applicable (for skips)</td>
<td>Counting by domain with expression of uncertainty (no adjustment)</td>
<td>Rate-based estimation with expression of uncertainty (no adjustment)</td>
<td>Enumeration-based estimation with expression of uncertainty (no adjustment)</td>
<td>Anchor on proxy and adjust</td>
<td>Search by domain</td>
<td>Based on cues used from question</td>
<td>“Never…”</td>
<td></td>
</tr>
<tr>
<td>Counting by observation</td>
<td>Rate-based estimation by domain</td>
<td>Rate-based estimation by domain</td>
<td>Rate-based estimation by domain</td>
<td>Based on another event</td>
<td>“Nowadays…/”usually”</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rate-based estimation by domain with adjustment for uncertainty</td>
<td>Rate-based estimation by domain with adjustment for uncertainty</td>
<td>Rate-based estimation by domain with adjustment for uncertainty</td>
<td>Based on regularity of behaviour</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
“Starts quarrels with others umm yeah I think definitely quarrels with my parents flash straight into my mind so agree a little probably I don’t like to say or admitting to bad things do you”

Participant 18, BFI item 12
RESPONSE: EDITING ANSWERS

It is well known that respondents do not like to give answers that they believe to be socially undesirable.

Sudman et al (1996) suggest asking respondents at the end of the questionnaire which items they found awkward to answer, they were embarrassed to say or felt were threatening.

I did not do this in my study, but I did find a few instances of this occurring:
GIVING SOCIALLY DESIRABLE RESPONSES

“Generates a lot of enthusiasm umm I try [chuckles] it’s a bit boastful to say that you make everyone enthusiastic around you agree strongly [chuckles] no I agree a little”
Participant 7, BFI item 16

“Starts quarrels with others umm yeah I think definitely quarrels with my parents flash straight into my mind so agree a little probably I don’t like to say or admitting to bad things do you”
Participant 18, BFI item 12
WHAT DO I DO WHEN I FIND A PROBLEM WITH A QUESTION?

1. Use Oskenberg et al’s arbitrary point of if 15% of your sample have a problem with an item. So make sure you identify all the different problems and how many are associated with each item.

2. Altering a question may make it worse, you cannot be sure therefore if you decide to alter the way a question is worded or presented, you must pilot again asking people to think aloud.

3. For retrieval, comprehension problems and judgement problems there is literature available suggesting which words, formats etc make it harder for people to respond in the way you want and ways that you can improve on this. Sudman et al’s book (1996) outlines a lot of these methods.
REFERENCES


REFERENCES


