

# Assessing inflation in variance estimates due to coder error using a coder reliability study

Patrick Sturgis  
Division of Social Statistics  
University of Southampton

# Overview

---

- Why ask 'open' questions in surveys?
- Types of coder error
- Implications of coder error for survey estimates
- Coder reliability studies
- Example - the 2000 UK Time Use Survey
- Conclusions and some further thoughts

please tell me, in your own words, what comes to mind when you think about the term “medical research?”

# Verbatims in surveys

## – pros and cons

---

### ■ Pros

- Increases the range of responses, particularly when *potential* range of responses unclear
- And where range of potential answers is vast (e.g. occupation)
- Allows respondent to interpret and respond to issue area in their own terms

### ■ Cons

- Still need to be converted to 'fixed' categories for analysis purposes
- This is error-prone
- ...and expensive

# Coder error

---

- Intuitive concern with coding ‘open’ responses relates to bias
- Coders applying the ‘wrong’ code
- But bias generally difficult/impossible to estimate for many variables
- What is the ‘correct’ code?
  - Compare to ‘expert’ coders?
  - Verification with respondent?

# Sources of coder error

---

- Lack of motivation, training, supervision, unsuitable environment etc.
- Complex frame, many codes
- Weaknesses in the frame, non-redundant and vague codes
- Technical problems (digital imaging, accessing online 'help' etc.)

# Components of Coder Error

---

- Coders apply different codes to same recorded response
- With respect to coders, such 'errors' can be either:
  - Uncorrelated (simple)
  - Correlated

# Uncorrelated (simple) coder error

---

- Coders apply different code to same answer
- But tendency to apply particular code not systematically associated with particular coder(s)
- Equivalent to Reliability (R) of code
- Measured by  $\bar{P}$ , % of all pair-wise comparisons that apply the same code to same response
- Not directly estimable from coded data (requires special studies)
- Reduces precision by factor of 1-R
- E.g. R=0.7 reduces effective sample size by 30%
- Conventional variance estimators unbiased

# Correlated Coder Error

---

- Coders apply different codes to same answer
- Tendency to select particular code(s) systematically higher for particular coder(s)
- Measured by rho (intra-class correlation)
- Conceptually equivalent to interviewer design effects
- Values of rho for coders have generally been found to be small, < 3% of total variance
- However, their impact can be large if average workloads are high and codes are unreliable

# Variance inflation due to correlated coder error

---

- Variance inflation factor (Biemer & Trewin, 1997):

$$ceff = (1 + p_c (m-1) / P_i)$$

- $P_c = \rho$ ;  $m =$  average workload;  $P_i =$  code reliability
- E.g.  $p_c = 0.02$ ;  $m = 1000$ ;  $P_i = 0.75$  results in variance inflation factor of 14.7
- $Ceft = \sqrt{ceff}$  = increase in standard error due to correlated coder error

# summary

---

- So, to correctly estimate variance of coded verbatim response, we require:
- Estimate of  $\rho$  for the code
- Estimate of reliability of the code
- Average size of workload of coders

---

Illustration:

The 2000 UK Time Use Survey  
(UKTUS)

# 2000 UKTUS Design

---

- Multi-stage probability sample of households
- All household members aged 8+ complete two 'own words' diaries
- Diaries coded to a hierarchical frame – 286 codes at lowest level, 10 at highest level
- E.g. 'Took Dog for a Walk'
  - 3 – Household and Family Care
    - 34 – Gardening and Pet Care
      - 344 – Walking the Dog

Morning Time, am	<b>What were you doing?</b> <i>Please record your main activity for each 10-minute period.</i>	<b>What else were you doing?</b> <i>Write in the most important activity you were doing at the same time</i>	<b>Where were you?</b>	<b>Were you with anybody?</b> <i>Please mark the boxes. See example on page 2.</i>				
	<i>Enter one main activity on each line.</i>	<i>e.g. Looking after children, listening to the radio or having a drink</i>	<i>e.g. At home, at friends, in car, on bus, train, cycling, walking</i>	Alone or with people you don't know	Children up to 9 living in your household	Children aged 10 to 14 living in your household	Other household members	Other persons that you know
7:00 - 7:10				<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7:10 - 7:20				<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7:20 - 7:30				<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7:30 - 7:40				<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7:40 - 7:50				<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7:50 - 8:00				<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
8:00 - 8:10				<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
8:10 - 8:20				<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
8:20 - 8:30				<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
8:30 - 8:40				<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
8:40 - 8:50				<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
8:50 - 9:00				<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
9:00 - 9:10				<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
9:10 - 9:20				<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
9:20 - 9:30				<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
9:30 - 9:40				<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
9:40 - 9:50				<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
9:50 - 10:00				<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>



05

PLEASE CONTINUE WHEN YOU HAVE COMPLETED ALL THE COLUMNS YOU NEED TO

000573



# Coder reliability study design

---

- Five coders coded the same 40 diaries
- Diaries selected purposively to ensure full coverage of the coding frame
- Few differences between activity distributions of study and full sample data
- Unit of analysis is the 'ten minute time slot'
- 1 day = 144 'ten minute time slots'
- So total activity codes =  $5 \times (40 \times 144) = 28,800$

# $\bar{P}$ for all codes by individual coders

Coder	3 digit level	1 digit level
Coder 1	91%	95%
Coder 2	90%	95%
Coder 3	90%	94%
Coder 4	88%	94%
Coder 5	88%	94%
<b>Overall</b>	<b>89%</b>	<b>94%</b>

*Note: 4/5 Coders in agreement,  $P = 60\%$*

# Reliabilities of Individual Codes

CODE	% OF ALL CODES	$\bar{P}$
PERSONAL CARE (V0)	43.65	97.2
EMPLOYMENT (V1)	6.44	95.7
STUDY (V2)	5.89	96.0
HOUSEHOLD/FAMILY CARE (V3)	15.47	91.2
VOLUNTARY WORK (V4)	2.77	87.4
SOCIAL LIFE (V5)	5.56	78.6
SPORTS/OUTDOOR ACTIVITIES (V6)	1.86	82.3
HOBBIES AND GAMES (V7)	2.58	89.9
MEDIA CONSUMPTION (V8)	9.07	94.4
TRAVEL AND UNSPECIFIED (V9)	6.71	86.3

# Variance Inflation factors

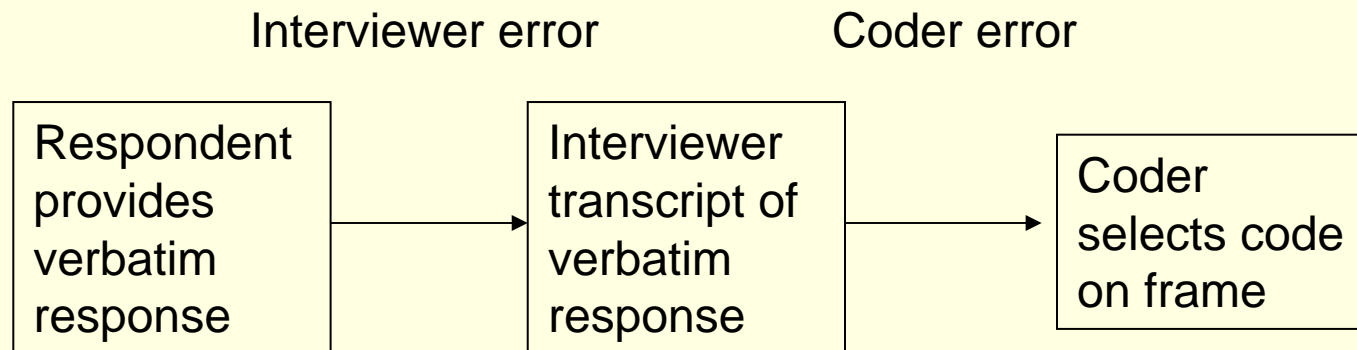
ACTIVITY CODE	rho	F	95% c.i.		$\bar{P}$	ceff	ceft
			Low	High			
PERSONAL CARE	0.004	12.0	0.0014	0.0381	97.2	1.34	1.16
EMPLOYMENT	0.002	2.3	0.000*	0.0368	95.7	1.26	1.12
STUDY	0.006	3.4	0.0004	0.0639	96.0	1.72	1.31
HOUSEHOLD/FAMILY CARE	0.005	6.1	0.0016	0.0559	91.2	2.58	1.61
VOLUNTARY WORK	0.000	3.6	0.000*	0.0206	87.4	1.00	1.00
SOCIAL LIFE	0.016	8.5	0.0043	0.1285	78.6	11.26	3.36
SPORTS/OUTDOOR ACTIVITIES	0.019	9.0	0.0016	0.1706	82.3	11.08	3.33
HOBBIES AND GAMES	0.034	4.5	0.0089	0.2484	89.9	11.29	3.36
MEDIA CONSUMPTION	0.013	9.5	0.0036	0.1071	94.4	3.17	1.78
TRAVEL AND UNSPECIFIED	0.016	5.9	0.0045	0.1296	86.3	3.00	1.73

Ceff and ceft assume average coder workloads of 3000.

# Some Examples

Code	Time	Point estimate	SRS S.E.	TRUE S.E.
PERSONAL CARE	08:20-08:30	52.4%	0.34	0.39
EMPLOYMENT	14:40-14:50	18.7%	0.27	0.30
STUDY	11:20-11:30	7.3%	0.18	0.24
HOUSEHOLD/FAMILY CARE	16:50-17:00	19.5%	0.27	0.43
VOLUNTARY WORK	15:30-15:40	2.1%	0.10	0.10
SOCIAL LIFE	22:30-22:40	3.5%	0.13	0.44
SPORTS/OUTDOOR ACTIVITIES	13:20-13:30	12.0%	0.22	0.73
HOBBIES AND GAMES	16:20-16:30	5.2%	0.15	0.50
MEDIA CONSUMPTION	21:20-21:30	43.4%	0.34	0.61
TRAVEL AND UNSPECIFIED	16:00-16:10	16.6%	0.26	0.45

# For CAPI this is only half the problem...



- 
- Illustration from a recent research project
  - Respondents asked why, in their own words, they had selected the middle response alternative on an attitude questions about EC and GM crops and foods
  - This is what I received...

# Example 'verbatim'

QuestionID	Responden	Verbatim	Codes	
tq6	1031	I DONT THINK THERE IS ENOUGH INFORMATION ABOUT GM FOODS TO GENERAL PUBLIC/		2
tq6	1141	GETTING TO THE TRUTH OF WHAT IS GOING ON IS LIKE GETTING BLOOD OUT OF STONE		2
tq6	115	i feel we should make our own rules and regulation, all this about human rights comes from there		2
tq6	1194	lack of knowledge		3
tq6	1225	i dont have enough information about them		3
tq6	1230	not hrd of it		3
tq6	1281	i have to a bit more research		1
tq6	1317	don't know enough about the workings of it tohave a view		3
tq6	1351	i simply dk enough		3
tq6	187	Don't like them as an institution but the people that are there probably do work hard		2
tq6	370		33	3
tq6	371		33	3
tq6	42	dontbuy		2
tq6	425	i don't think i have enough information to make a decision //		3
tq6	440	largely irrelevent		2
tq6	442	Economics under discussion re \"meltdown\"		2
tq6	50		3	3
tq6	568	i dont know a great deal about it		3
tq6	606	II DO NOT HAVE ENOUGH INFO ON THE SUBJECT/		3
tq6	732	NOT ENOUGH KNOWLEDGE		3
tq6	876	er		3
tq6	877	er		3
tq6	924	it is not clear what there function is		2
tq6	977	dont know much about it		3
tq6	1599	WE HAVE ENTERED IT AND WLD COST ALOT TO CME OUT/-		2
tq6	1775	I DO NOT KNOW ENOUGH ABOUT THEIR ROLE		3
tq6	1914	THE WAY THEY DO IT/ TAKE TOO LONG TO DO ANY THING FOR THE GENERAL PUBLIC/		2
tq6	1985	I DON'T KNOW ENOUGH ABOUT WHAT THEY ARE DOING.		3
tq6	2072	EVIDENCEONE WAY OR OTHER/TOO EARLY ATMOMENT		1
tq6	2102	more evidence is required		1
tq6	2215	not convinced of the evidence either way		1

# Conclusions

---

- Conventional variance estimators (heroically) underestimate variance of coded verbatims
- This is partly because they tend to use a small number of coders
- e.g. increasing from 7 to 14 coders on UK 2000 would reduce s.e. inflation factor by a third
- But may also reflect lack of quality monitoring of code frames and coders
- Gains from more coders must be balanced against complexity and cost of hiring additional coders
- We have only considered the coder error, transcription error is likely to make matters considerably worse

# Further thoughts

---

- Continuous monitoring should be implemented to identify weaknesses in frame
- and amongst individual coders
- Coder reliability studies implemented as standard?
- Greater transparency of coding procedures in survey documentation
- Given the errors and associated costs, need to think hard about the potential benefits of verbatim questions