

NatCen

Social Research that works for society

Innovating with online data collection

Measurement, sampling and new
forms of data





Introduction

Online data collection has created new challenges & opportunities, and re-cast old ones

Will look at three pieces of methodological work on the NatCen Panel, relating to measurement, sample quality, and new forms of data:

- 'Don't know' answers in a web survey
- Implementing a 'Targeted design' on a web panel
- Linking survey and Twitter data



The NatCen Panel

- First probability-based research panel in GB open to be used by the social research community
- c.8,000 members recruited from face-to-face probability-based BSA survey
- Sequential mixed-mode fieldwork design (web/CATI), lasting c. one month
- Aims to produce reliable estimates for the British population in a shorter time-frame and at a lower cost than 'traditional' probability-based approaches.

‘Don’t know’ answers in web surveys





Background

- ‘Don’t know’ answer options typically not presented up-front
 - Concern that satisficing will lead to loss of data
- In interviewer-administered modes, rely on ‘spontaneous’ answers – not possible in web mode
- In web, DKs ‘hidden’, then appear if question left blank, but...
 - Mis-coding/exit the survey if aren’t aware of skipping option
 - Comparability to other modes...?



Experimental design

DK only shown when skipped	DK shown upfront
1. No explanation of functionality	3. No probe/prompt
2. Explanation	4. Polite probe/prompt when DK selected



Hidden (1)

To what extent do you agree or disagree with the following statement:

Immigrants are generally good for Britain's economy.

- Strongly agree
- Agree
- Neither agree or disagree
- Disagree
- Strongly disagree



Hidden (2)

Please provide an answer or select “Don’t know” or “Prefer not to answer”

To what extent do you agree or disagree with the following statement:

Immigrants are generally good for Britain’s economy.

- Strongly agree
- Agree
- Neither agree or disagree
- Disagree
- Strongly disagree
- Don’t know
- Prefer not to answer



Explanation

Before we start, we'd like to remind you that even if you are unsure of your answers, we still value your opinion. However, you do not have to answer any question you do not want to.

If we ask a question you don't know the answer to, or you would prefer not to answer, simply leave the question blank and click the 'Next' button to make the options 'Don't know' and 'Prefer not to say' appear.



How interested would you say you are in health research?

- Very interested
- Fairly interested
- Not very interested
- Not at all interested

◀ Previous

Next ▶



Please provide an answer or select "Don't know" or "Prefer not to answer"

How interested would you say you are in health research?

- Very interested
- Fairly interested
- Not very interested
- Not at all interested

- Don't know
- Prefer not to answer

◀ Previous

Next ▶



Upfront

To what extent do you agree or disagree with the following statement:

Immigrants are generally good for Britain's economy.

- Strongly agree
- Agree
- Neither agree or disagree
- Disagree
- Strongly disagree

- Don't know



Upfront + follow-up probe

Thank you for your response.

We would really like to know your opinion on this, even if you are unsure or don't feel you know enough about the topic.

If you had to choose, to what extent do you agree or disagree with the following statement:

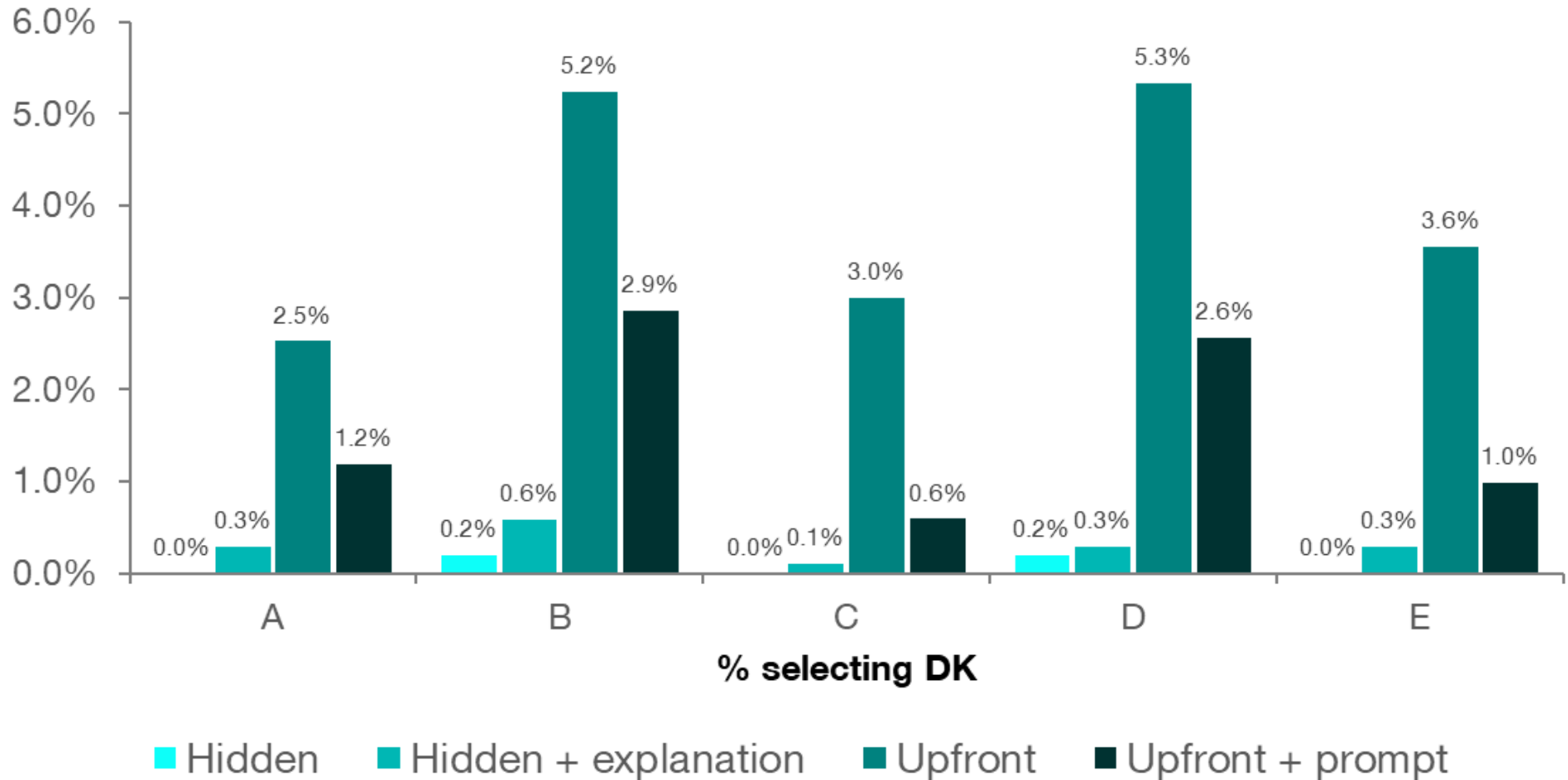
Immigrants are generally good for Britain's economy.

- Strongly agree
- Agree
- Neither agree or disagree
- Disagree
- Strongly disagree

- I really don't know



Proportion of DKs



Base: web only, hidden (1038), hidden + explanation (1025), upfront (1070), upfront + prompt (1014)

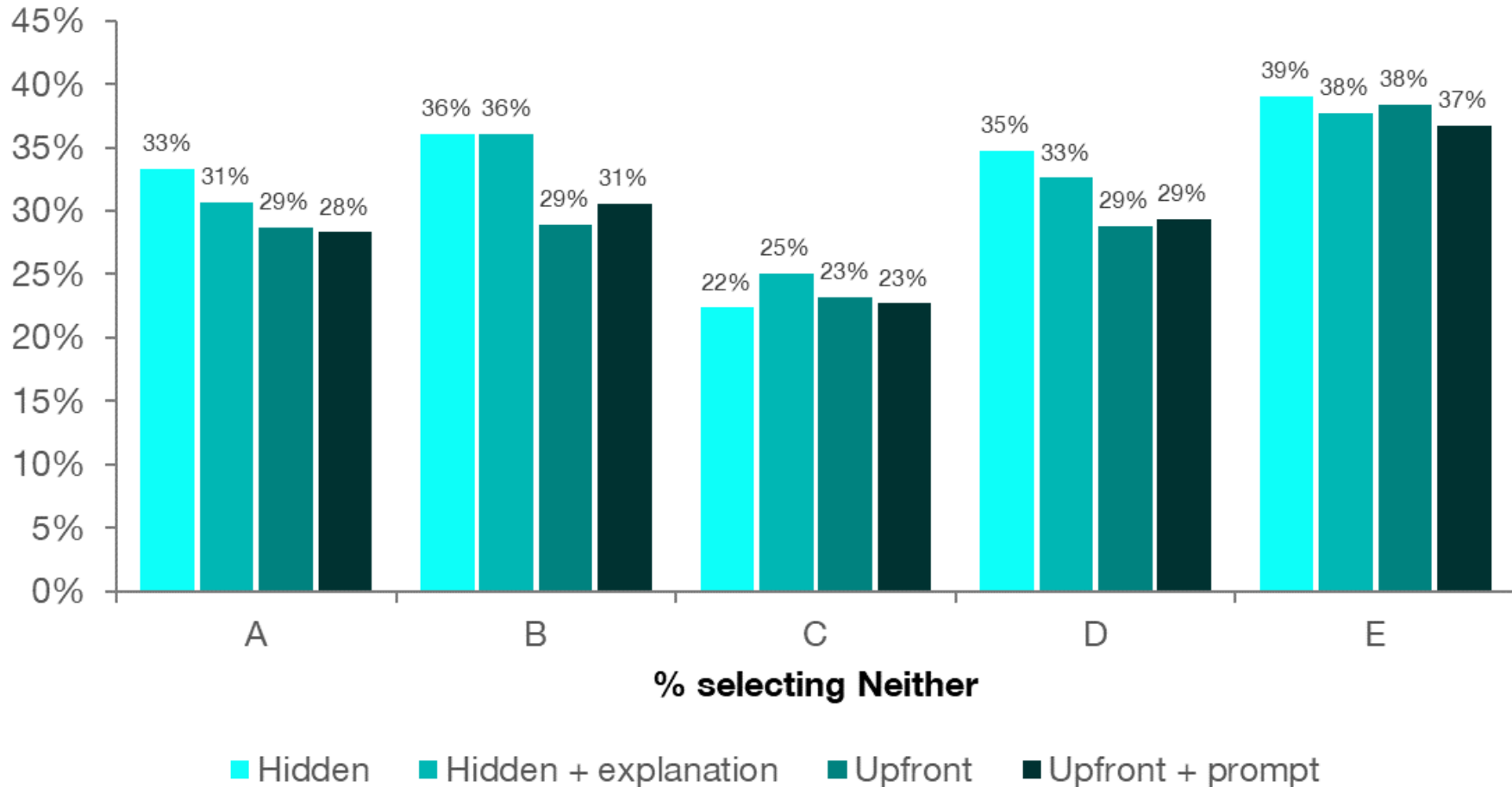


Impact of probes

- For those presented DK up-front, 3% of the probed selected it more than once, compared to 4% of those not probed (not statistically significant)
- On average, 42% of those probed still said 'DK'
- When asked why they coded 'DK':
 - 58% 'don't know enough about the topic' (42% of un-probed)
 - 18% 'my answer would vary depending on the situation' (33% of un-probed)



Effect on answer distributions



Base: web only, hidden (1038), hidden + explanation (1025), upfront (1070), upfront + prompt (1014)



Discussion points

- Explaining the 'DK' functionality makes little impact on use
- Presenting 'DK' up-front increases its use
- Probing DKs 'converts' many of these to substantive answers
 - Some indication of better *quality* answers?
 - Additional analytical utility?
- But...
 - Additional programming/data management costs?
 - Frustration of being asked the same question twice?
 - Impact on other types of question?
 - Does it improve comparability to other modes?

Implementing a 'Targeted design' on a web panel



2



Background

- Indications of (gradually) declining response rates and (gradually) increasing DEFFs from weighting on the Panel in mid-2017
- Addressed by ‘refreshment’ sample from new BSA waves, but wanted to minimise effect & considerations of longer-term studies
- Decided to implement a ‘Targeted design’, with goal of **improving the sample profile, while keeping costs, fieldwork length, and response rates neutral**



What is a 'Targeted design'?

- Form of 'adaptive' or 'responsive' design
 - Use of auxiliary data to target fieldwork protocols to sub-groups within the sample, with the goal of improving fieldwork outcomes
- Specifically, uses data available ahead of fieldwork start to group participants into different fieldwork designs



Prioritising cases

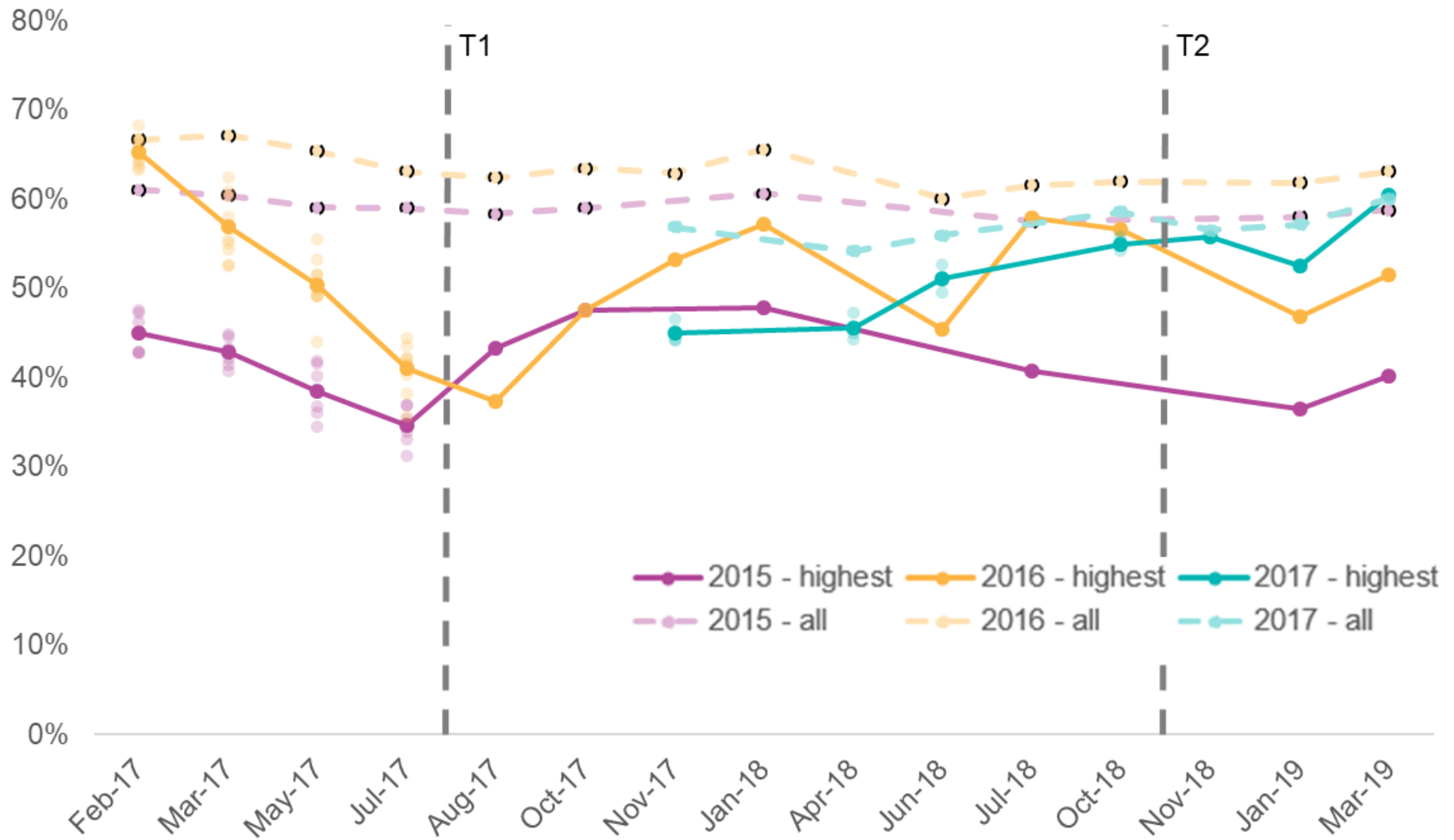
	Participated in all waves	Participated in some waves	Participated in no waves
1 (most under-represented)	Medium priority	Highest priority	Low priority
2	Medium priority	High priority	Low priority
3	Medium priority	High priority	Low priority
4	Medium priority	High priority	Low priority
5	Low priority	Medium priority	Lowest priority
6	Low priority	Medium priority	Lowest priority
7	Low priority	Medium priority	Lowest priority
8 (most over-represented)	Low priority	Medium priority	Lowest priority



Targeted fieldwork protocols

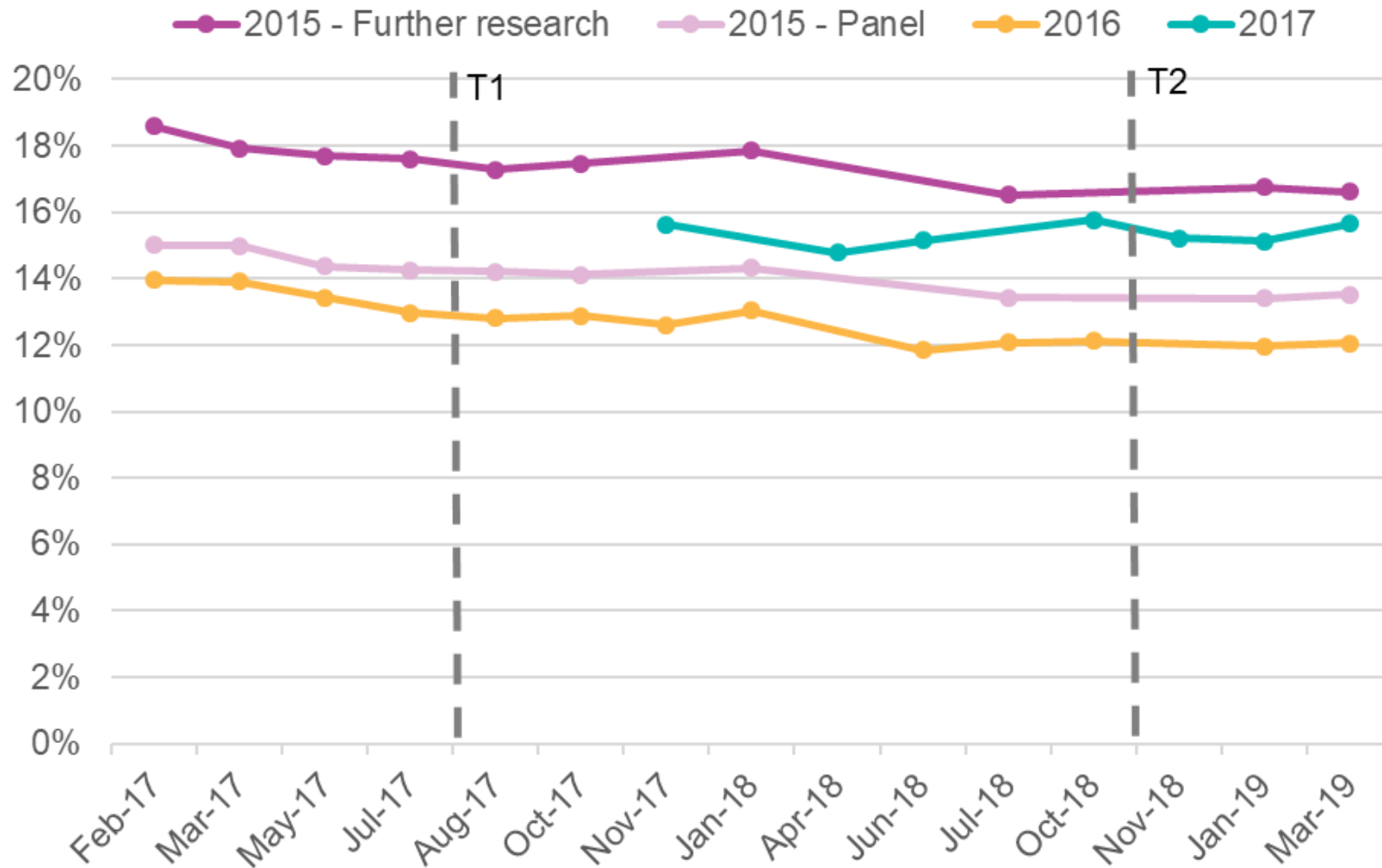
	Incentive offer	CATI fieldwork	Comms
Highest priority	£10	Min. of 8 calls	Two reminder letters
High priority	£5	Min. of 8 calls	One reminder letter
Medium priority	£5	Min. of 6 calls	One reminder letter
Low priority	£5	Min, of 4 calls	No reminder letters
Lowest priority	£5	Not issued	No reminder letters

Impact: Survey response rates of priority groups



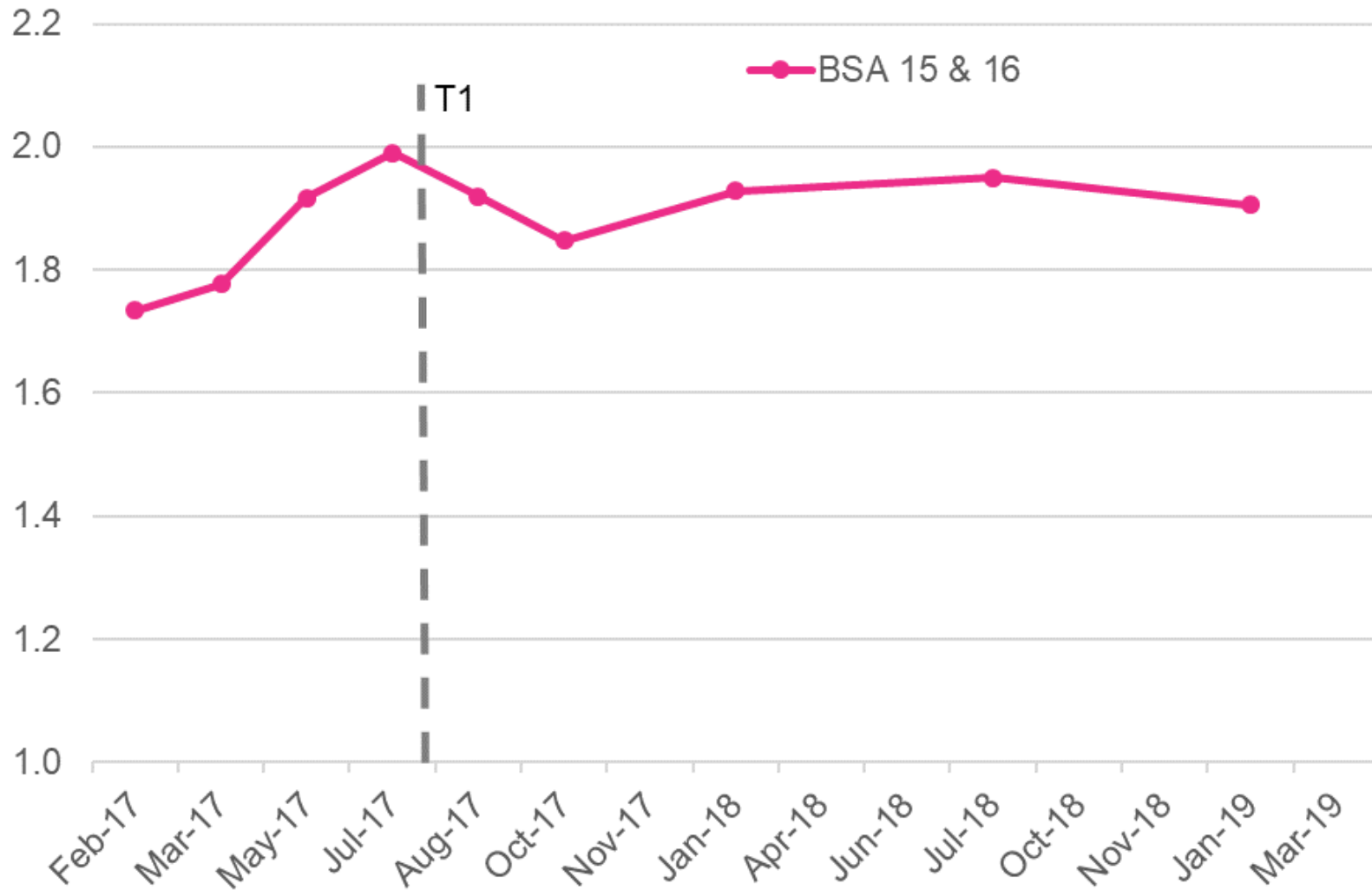


Impact: Overall response rates



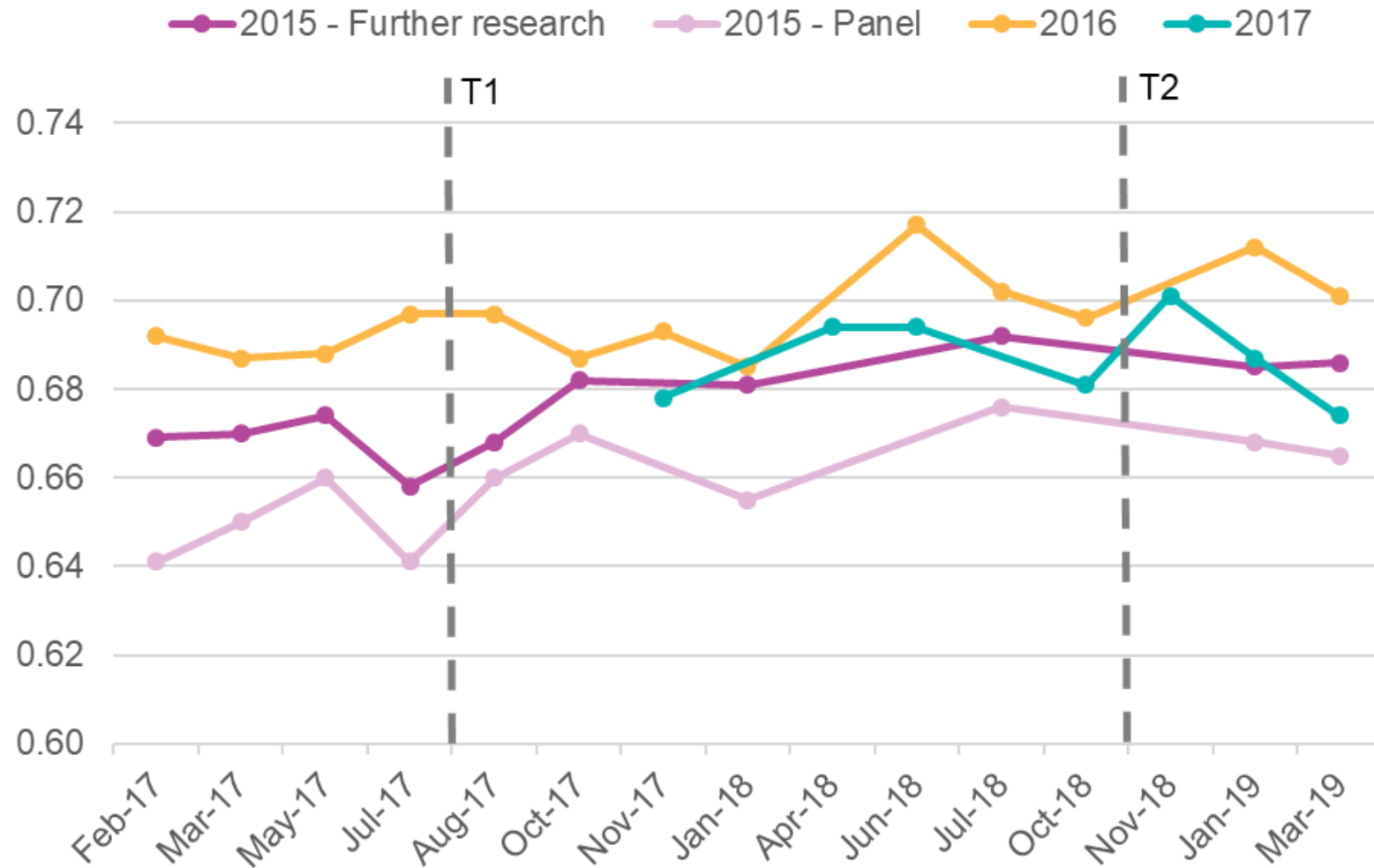


Impact: DEFFs





Impact: R-Indicators





Discussion points

- It is feasible to implement a targeted design
- Difficult to see any impact
 - Groups are too small?
 - Protocol changes too small or too late (for established panel sample)?
 - Variation between waves disguise overall trends?
 - It's not working?
- Measures are inconsistent
 - DEFF, response rates, R-indicators don't correlate as expected

But will probably continue...

Linking survey and Twitter data

3



Why?

- Continued & increasing interest in using social media data for social research
 - Social media as a phenomenon to be studied
 - Social media as a lens through which to view the general population
- But they continue to have their drawbacks
- Linking survey & social media data attempts to address some of these by:
 - Collecting informed consent
 - Putting sample in context of the population
 - Understanding whose data you are analysing
 - Validating machine-based classifications



Consent

- As we are in contact with participants, have the opportunity to ask people for consent to access their Twitter data (and link it to their survey answers)
- Need to balance providing enough information, with enough detail that consent is informed, but not so much that people are overwhelmed and don't read it (yet alone take it in)



Consent – Question wording

As social media plays an increasing role in society, we would like to know who uses Twitter, and how people use it. We are also interested in being able to add people's, and specifically your, answers to this survey to publicly available information from your Twitter account such as your profile information, tweets in the past and in future, and information about how you use your account.

Your Twitter information will be treated as confidential and given the same protections as your interview data. Your Twitter username, and any information that would allow you to be identified, will not be published without your explicit permission.



Consent - Helplinks

What information will you collect from my Twitter account?

What will the information be used for?

Who will be able to access the information?

What will you do to keep my information safe?

What if I change my mind?



Consent rates

	Web	CAT/F2F	Overall
BSA 2015	-	37%	37%
NatCen Panel	26%	34%	27%
USoc IP10	24%	41%	31%



Maintaining security

- Cannot rely on ‘anonymisation’ to maintain data security
- But there are other approaches to help minimise risks:
 - Systematic processing
 - Data reduction
 - Controlled access
 - Data deletion



Discussion points

Methodological

- Linking is feasible – need to demonstrate its value
- Key challenge of low consent rates
- Can it be expanded to other platforms? Is Twitter ‘the future’?

Ethics

- Social media data disrupt established approaches
 - Return to principles, rather than default ‘tools’
- On-going discussions of best approaches

Fin

Curtis Jessop

Research Director

Curtis.Jessop@natcen.ac.uk

@curtisjessop



NatCen

Social Research that works for society