Linking survey and social media data: Experiences and evidence

Presentation to the University of Michigan.

Tarek Al Baghal
Deputy Director, Understanding Society
University of Essex

1

Introduction

Deputy Director, Understanding Society

- UK Household Longitudinal Study, among largest in the world
- Lead for enhancements and questionnaire development

Professor of Survey Methodology, Institute for Social and Economic Research

UK Parliamentary Academic Fellow

Fellow of Higher Education Academy

- Questionnaire Design
- Applied Sampling
- Longitudinal Survey Design and Analysis

2

Research

Focus on linkages between surveys and new forms of data

- Sea Hero Quest spatial cognition app (game)
- Biomarkers
  - Mode differences (including nurses)
  - Blood/microbiome
- **Social media/survey data linkage**

3

Acknowledgments

ESRC grant: “Understanding [Offline/Online] Society: Linking Surveys with Twitter Data”

- Luke Sloan – University of Cardiff
- Curtis Jessop – NatCen for Social Research
- Matthew Williams – University of Cardiff

National Centre for Social Research

CARDIFF UNIVERSITY

University of Essex

4

1

Background

5

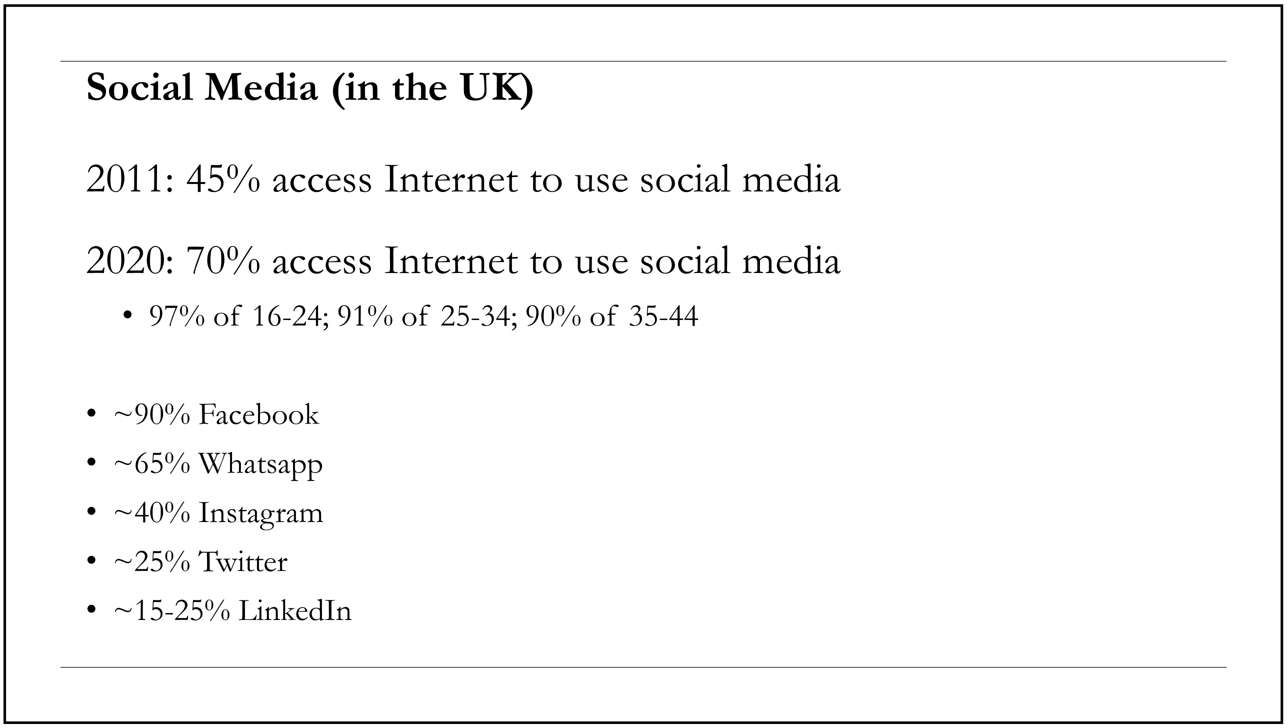
5

What are we trying to do, and why?

- Link survey participants' answers to publicly available information from their Twitter accounts
- Allows survey data to benefit from real-time, 'natural' behavioural and attitudinal data
- Adds the 'who' to Twitter data – creates a sample frame, and allows for the analysis of different groups
- Complement, not contrast

6

6



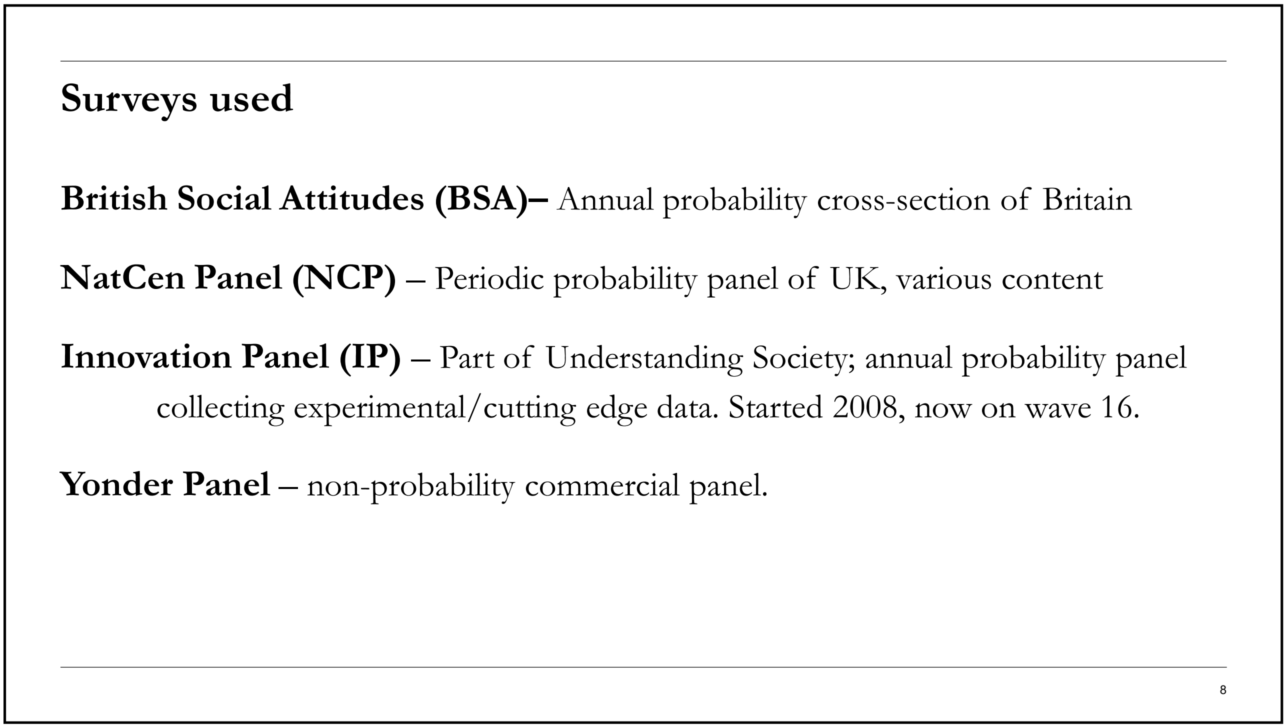
Social Media (in the UK)

2011: 45% access Internet to use social media

2020: 70% access Internet to use social media

- 97% of 16-24; 91% of 25-34; 90% of 35-44

- ~90% Facebook
- ~65% Whatsapp
- ~40% Instagram
- ~25% Twitter
- ~15-25% LinkedIn



Surveys used

British Social Attitudes (BSA)– Annual probability cross-section of Britain

NatCen Panel (NCP) – Periodic probability panel of UK, various content

Innovation Panel (IP) – Part of Understanding Society; annual probability panel collecting experimental/cutting edge data. Started 2008, now on wave 16.

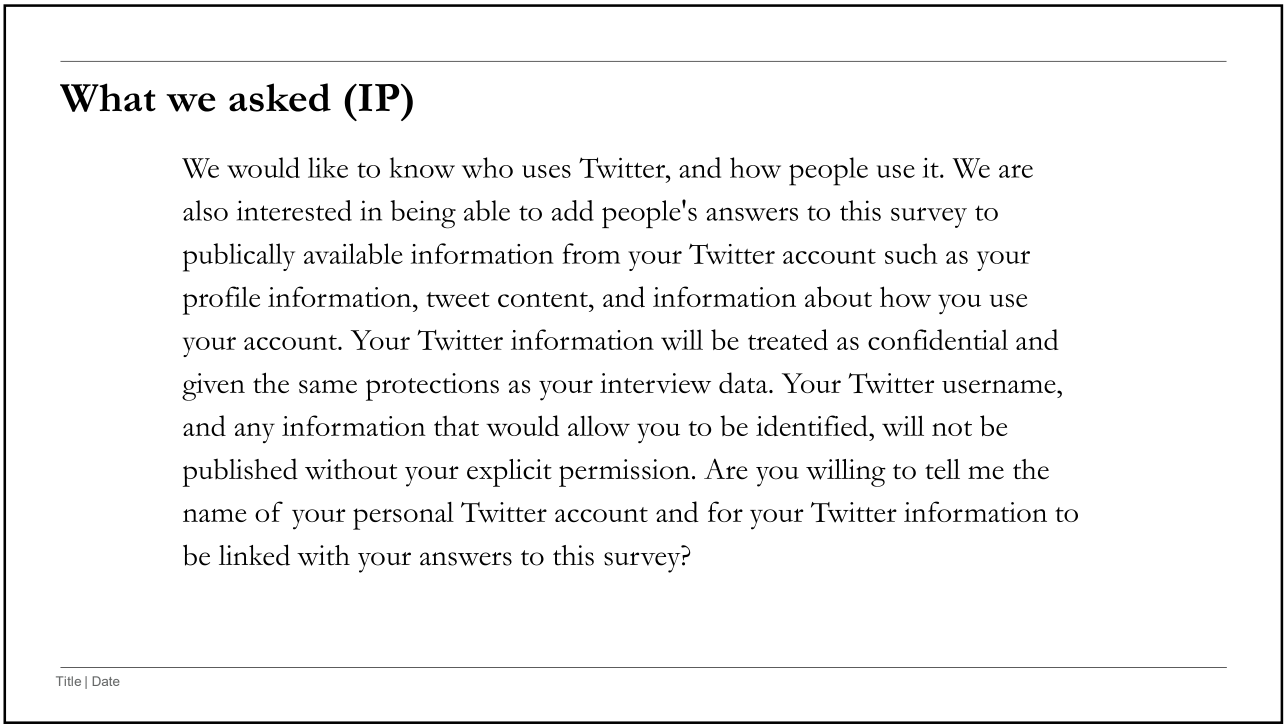
Yonder Panel – non-probability commercial panel.

2

Consent to link survey and social media data - initial evidence

Data collection

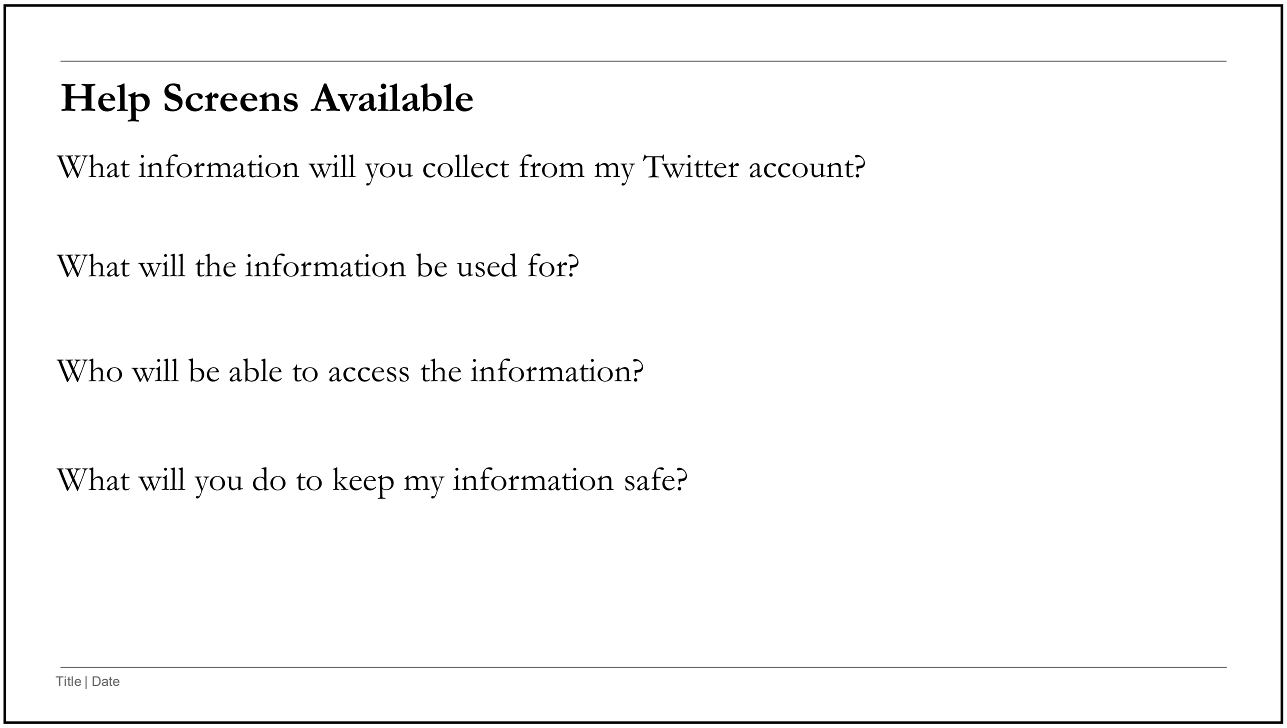
	BSA 2015	NCP 2017	IP10
Twitter users (n)	791	558	428
Mode	F2F	Web/Tel	F2F/Web
Fieldwork dates	Aug-Oct 2015	Jul 2017	May-Nov 2017
Incentive	£5	£5	£10-£30
Sample type	Probability cross-sectional	Probability panel, Based on BSA sample	Probability panel



What we asked (IP)

We would like to know who uses Twitter, and how people use it. We are also interested in being able to add people's answers to this survey to publically available information from your Twitter account such as your profile information, tweet content, 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. Are you willing to tell me the name of your personal Twitter account and for your Twitter information to be linked with your answers to this survey?

11



Help Screens Available

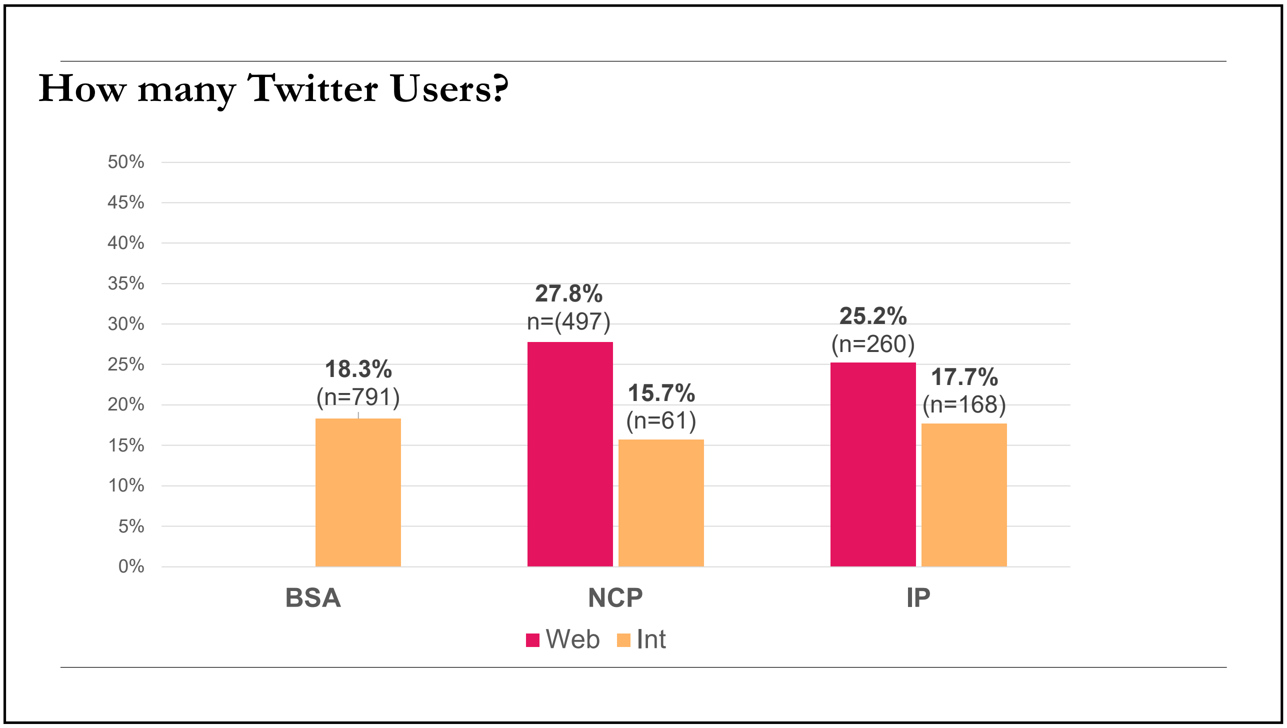
What information will you collect from my Twitter account?

What will the information be used for?

Who will be able to access the information?

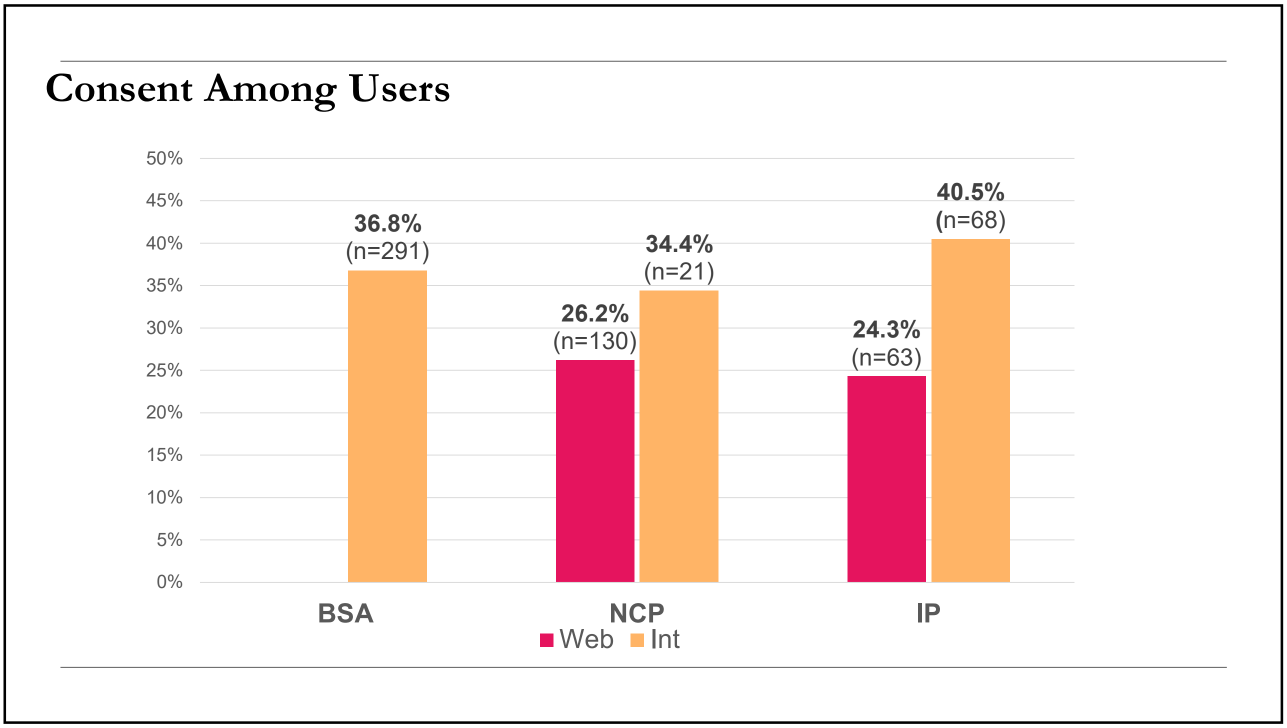
What will you do to keep my information safe?

12



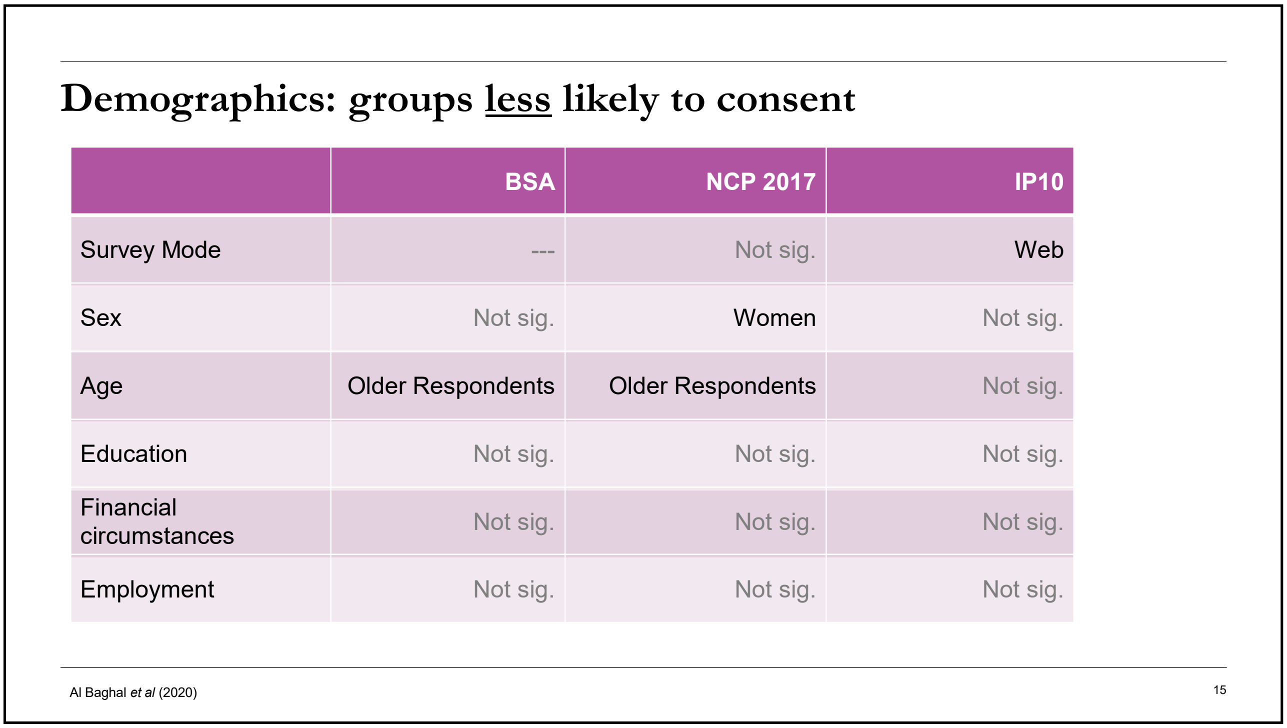
How many Twitter Users?

13



Consent Among Users

14



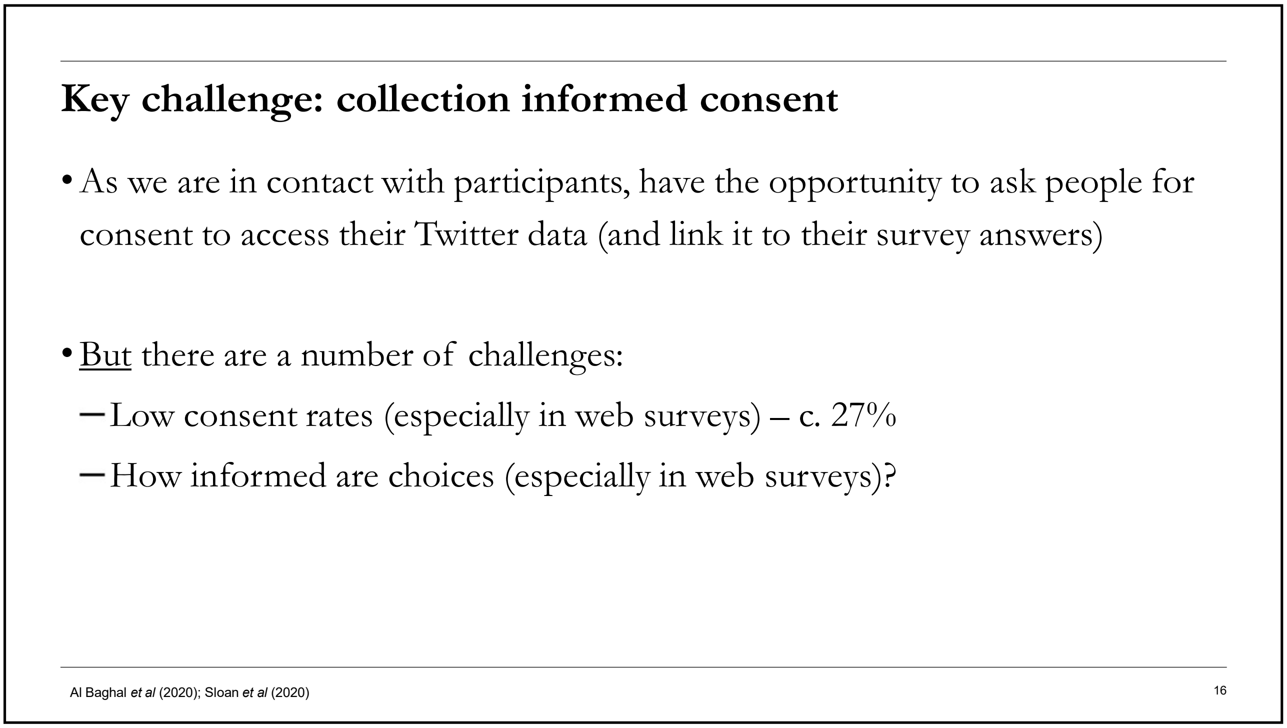
Demographics: groups less likely to consent

	BSA	NCP 2017	IP10
Survey Mode	---	Not sig.	Web
Sex	Not sig.	Women	Not sig.
Age	Older Respondents	Older Respondents	Not sig.
Education	Not sig.	Not sig.	Not sig.
Financial circumstances	Not sig.	Not sig.	Not sig.
Employment	Not sig.	Not sig.	Not sig.

Al Baghal *et al* (2020)

15

15



Key challenge: collection informed consent

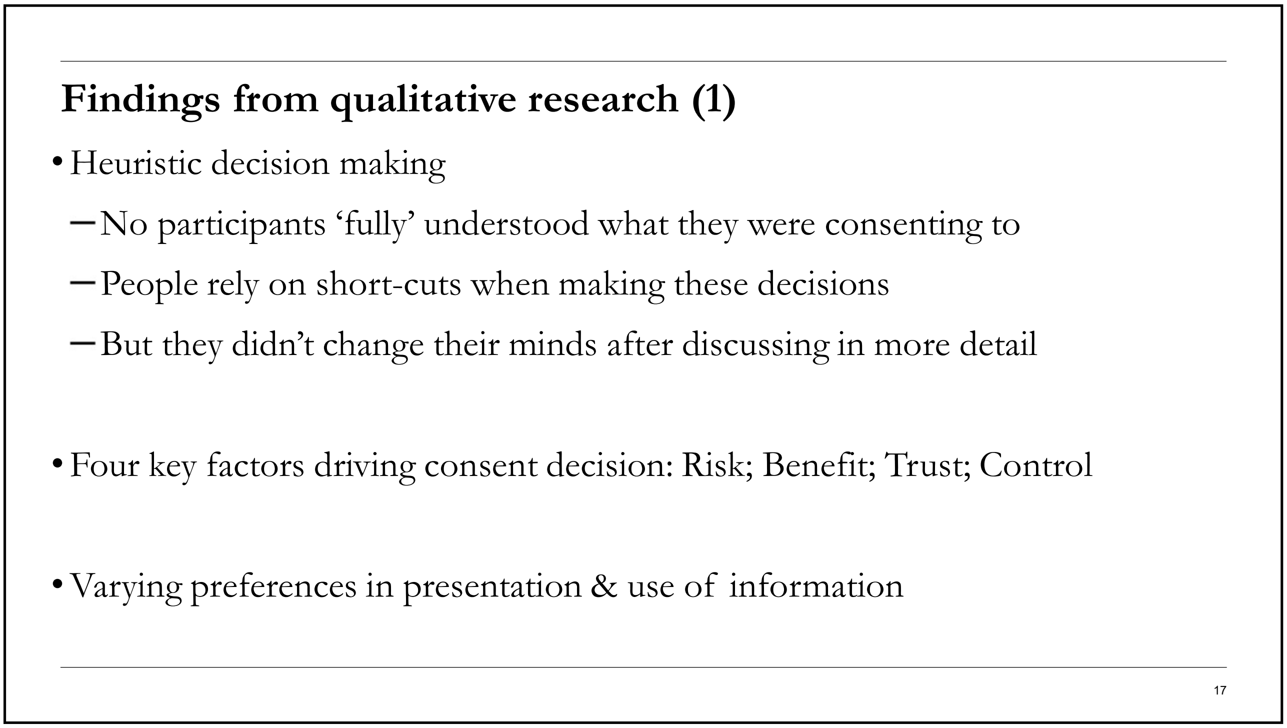
- As we are in contact with participants, have the opportunity to ask people for consent to access their Twitter data (and link it to their survey answers)

- But there are a number of challenges:
  - Low consent rates (especially in web surveys) – c. 27%
  - How informed are choices (especially in web surveys)?

Al Baghal *et al* (2020); Sloan *et al* (2020)

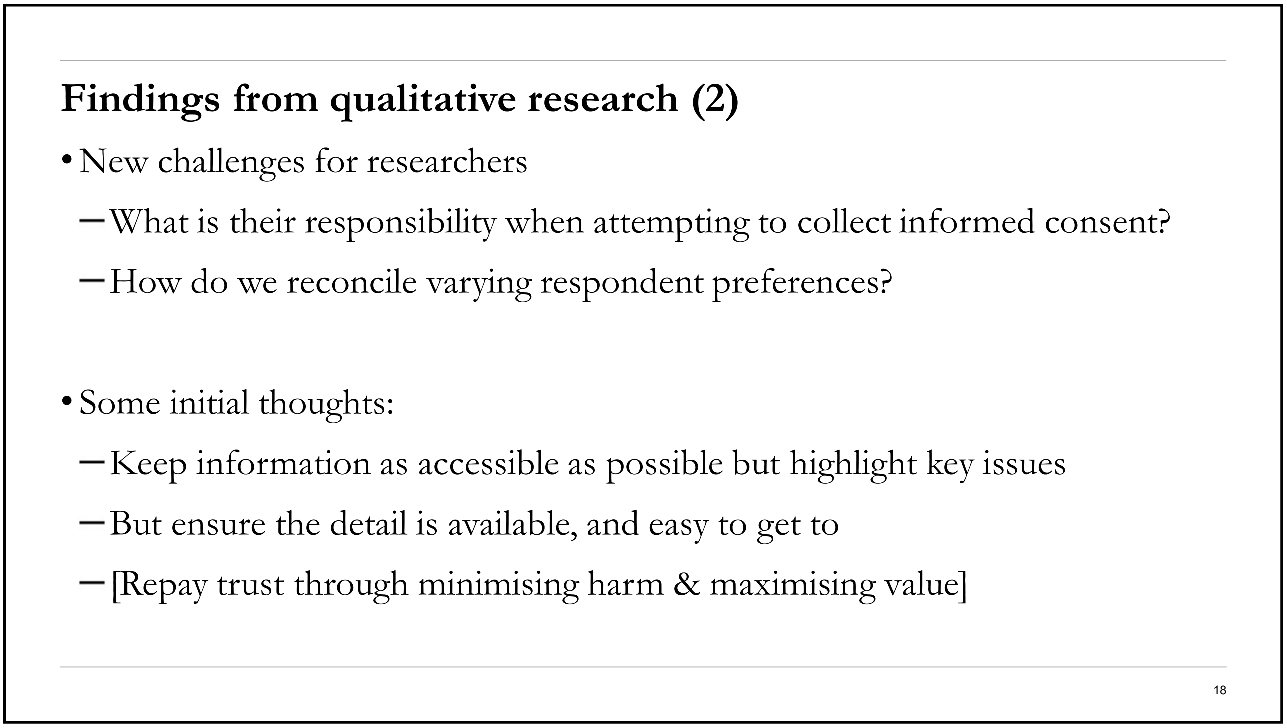
16

16



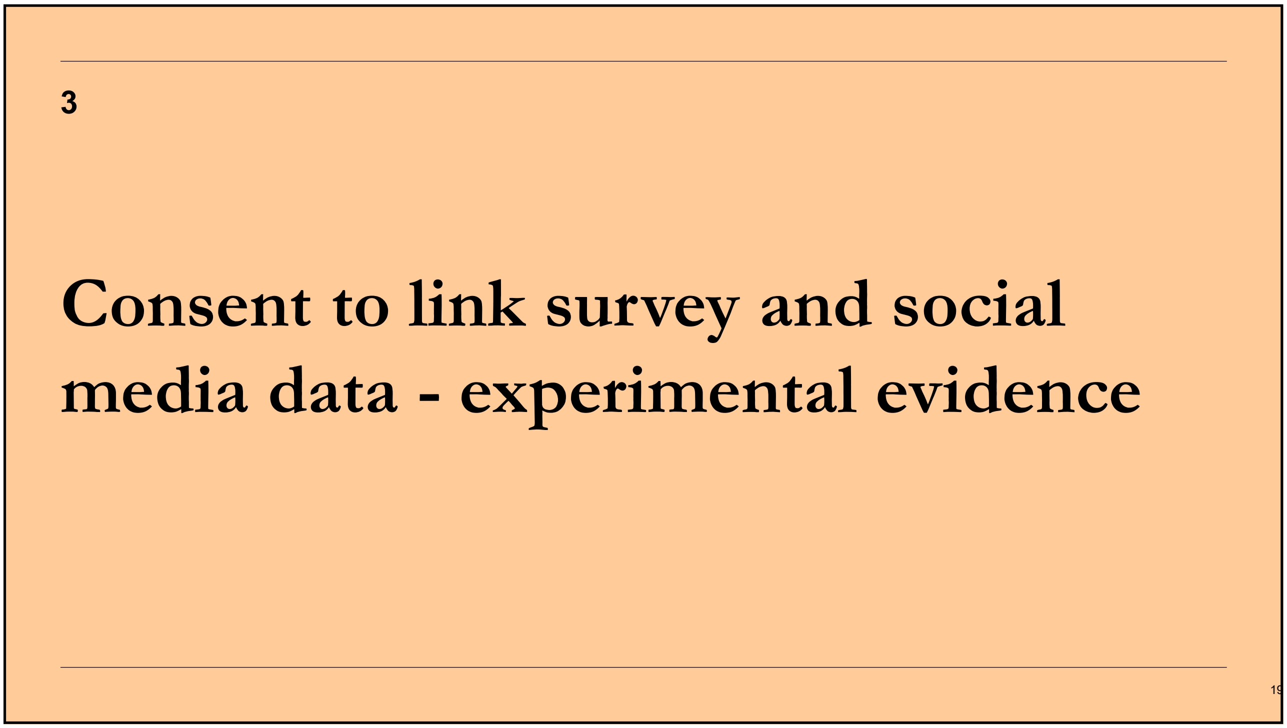
Findings from qualitative research (1)

- Heuristic decision making
  - No participants ‘fully’ understood what they were consenting to
  - People rely on short-cuts when making these decisions
  - But they didn’t change their minds after discussing in more detail
- Four key factors driving consent decision: Risk; Benefit; Trust; Control
- Varying preferences in presentation & use of information



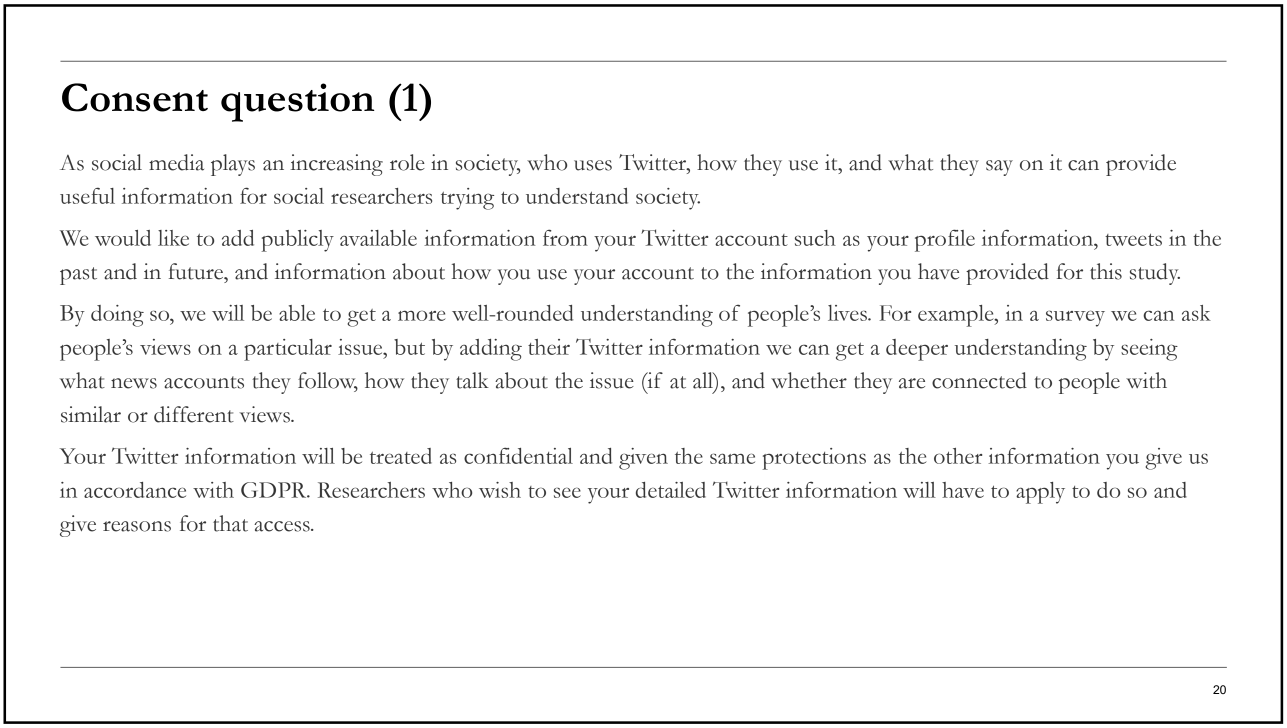
Findings from qualitative research (2)

- New challenges for researchers
  - What is their responsibility when attempting to collect informed consent?
  - How do we reconcile varying respondent preferences?
- Some initial thoughts:
  - Keep information as accessible as possible but highlight key issues
  - But ensure the detail is available, and easy to get to
  - [Repay trust through minimising harm & maximising value]



3

Consent to link survey and social media data - experimental evidence



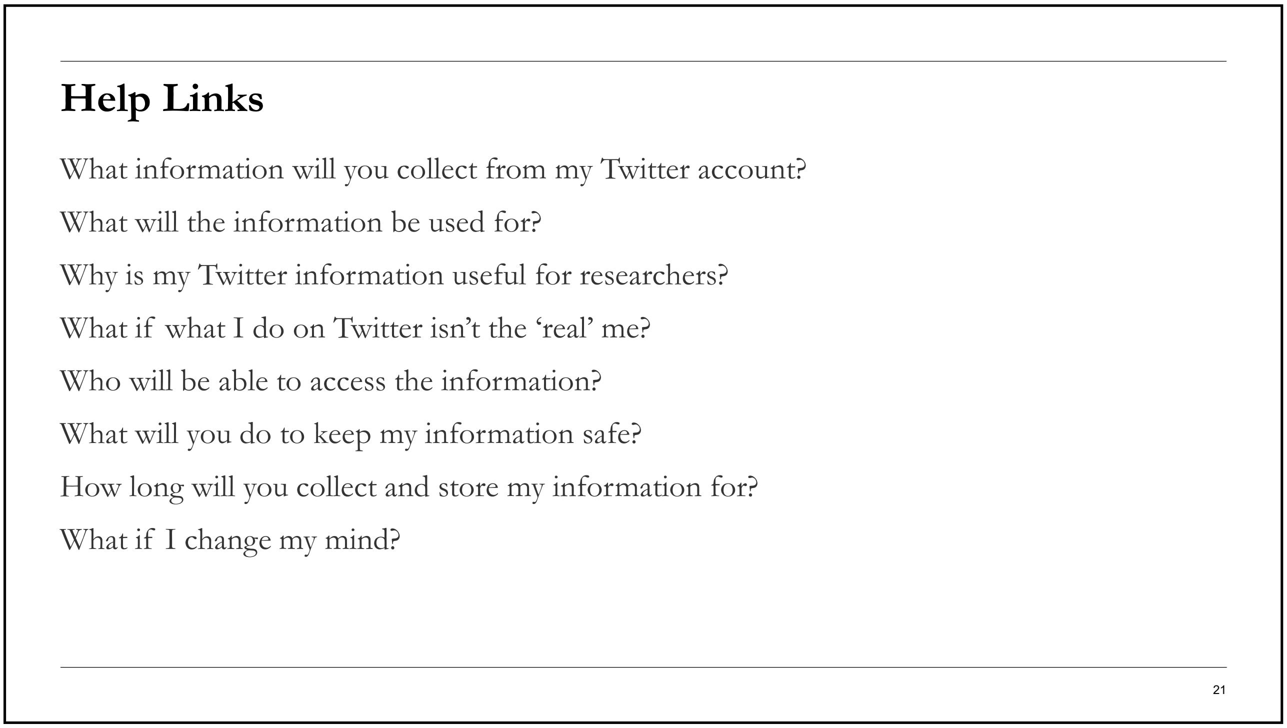
Consent question (1)

As social media plays an increasing role in society, who uses Twitter, how they use it, and what they say on it can provide useful information for social researchers trying to understand society.

We would like to add 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 to the information you have provided for this study.

By doing so, we will be able to get a more well-rounded understanding of people's lives. For example, in a survey we can ask people's views on a particular issue, but by adding their Twitter information we can get a deeper understanding by seeing what news accounts they follow, how they talk about the issue (if at all), and whether they are connected to people with similar or different views.

Your Twitter information will be treated as confidential and given the same protections as the other information you give us in accordance with GDPR. Researchers who wish to see your detailed Twitter information will have to apply to do so and give reasons for that access.



Help Links

What information will you collect from my Twitter account?

What will the information be used for?

Why is my Twitter information useful for researchers?

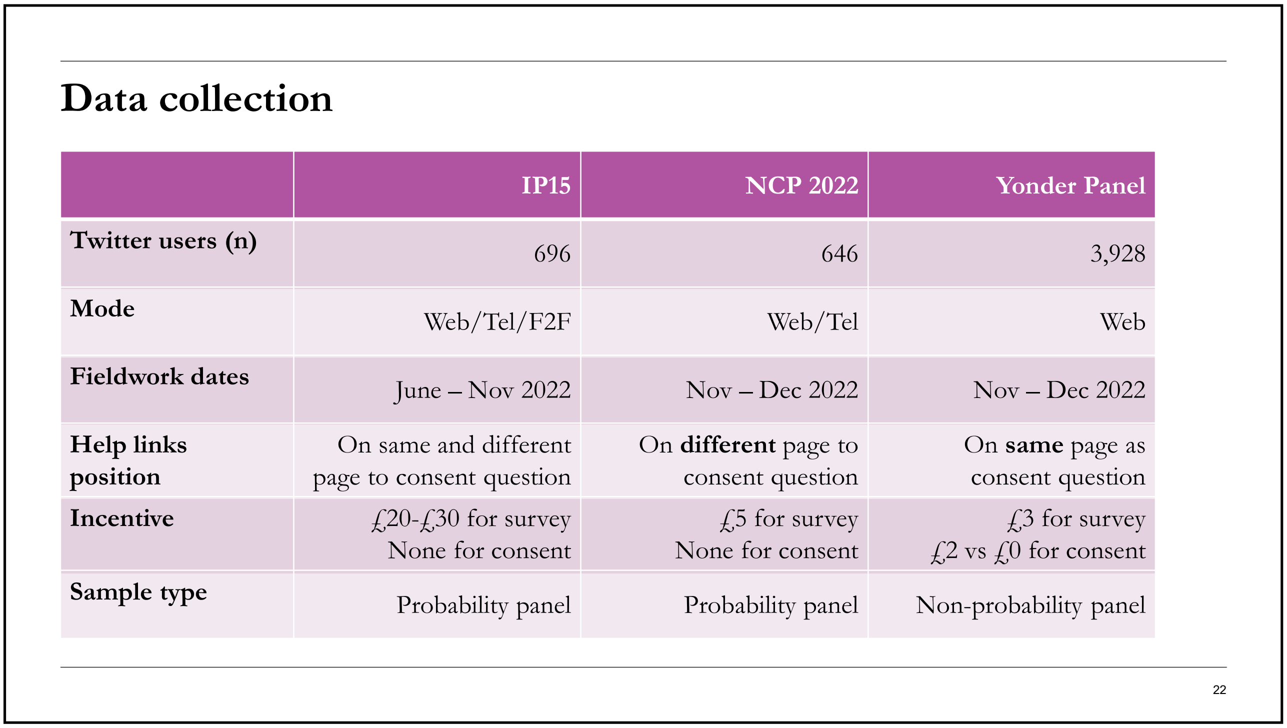
What if what I do on Twitter isn't the 'real' me?

Who will be able to access the information?

What will you do to keep my information safe?

How long will you collect and store my information for?

What if I change my mind?



Data collection

	IP15	NCP 2022	Yonder Panel
Twitter users (n)	696	646	3,928
Mode	Web/Tel/F2F	Web/Tel	Web
Fieldwork dates	June – Nov 2022	Nov – Dec 2022	Nov – Dec 2022
Help links position	On same and different page to consent question	On **different** page to consent question	On **same** page as consent question
Incentive	£20-£30 for survey None for consent	£5 for survey None for consent	£3 for survey £2 vs £0 for consent
Sample type	Probability panel	Probability panel	Non-probability panel

Experiment with help link positioning

Are you willing to tell us the username for your personal Twitter account, and for your Twitter information to be collected and added to the information you have provided for this study?

Group 1:

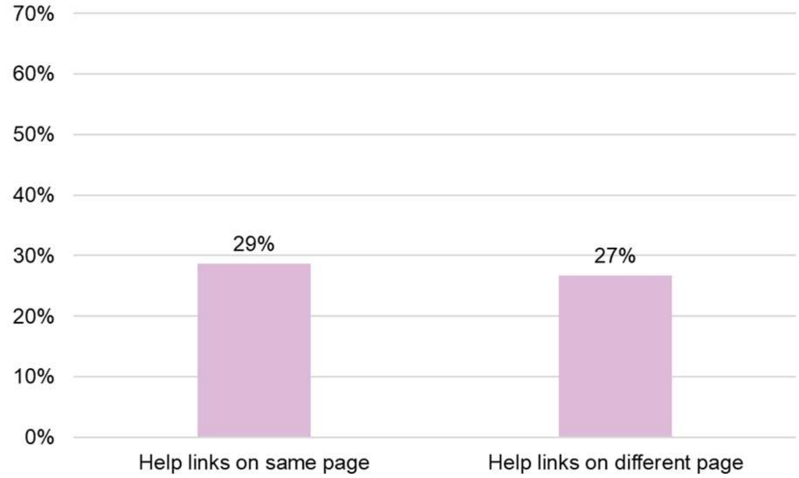
[HELPLINKS PRESENTED UP-FRONT]

1. Yes
2. No

Group 2:

1. Not sure, I would like more information [GO TO HELPLINKS PAGE]
2. Yes
3. No

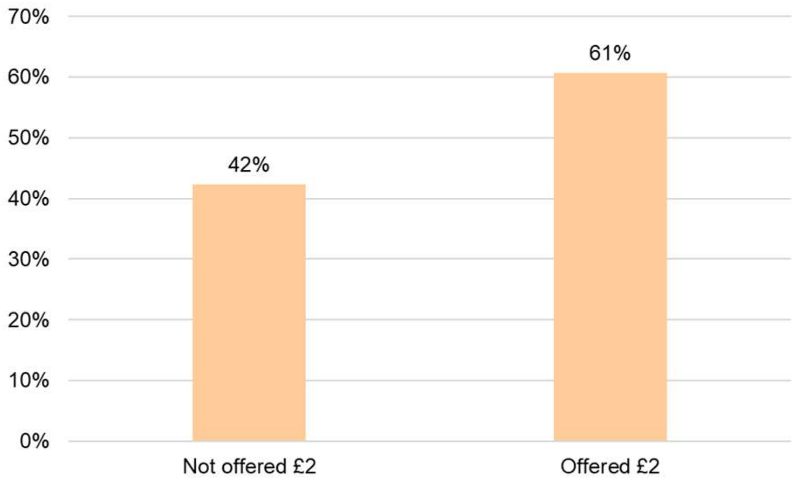
Consent to link survey & Twitter data by presentation of additional information



- 4% (n=14) participants asked to see more information

Source: IP15; Base: GB adults (16+) with a Twitter account: Help links on same page (356); Help links on different page (340)

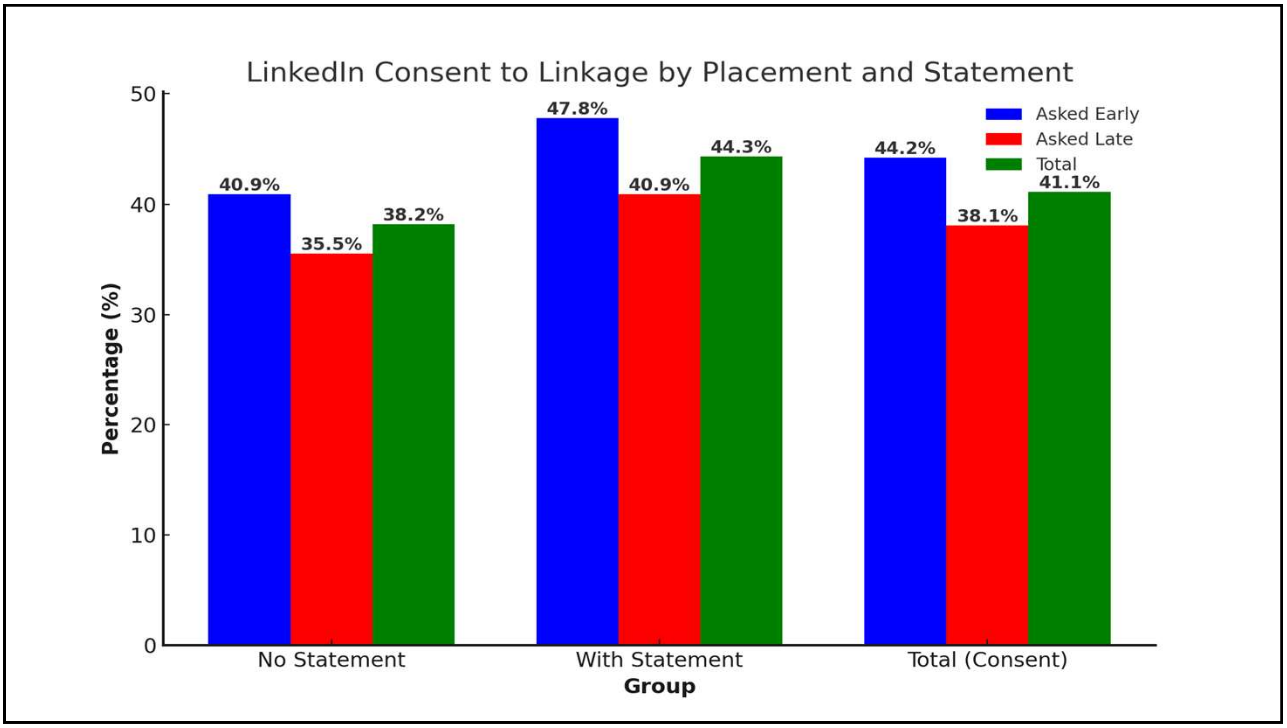
Consent to link survey & Twitter data by whether offered £2 incentive



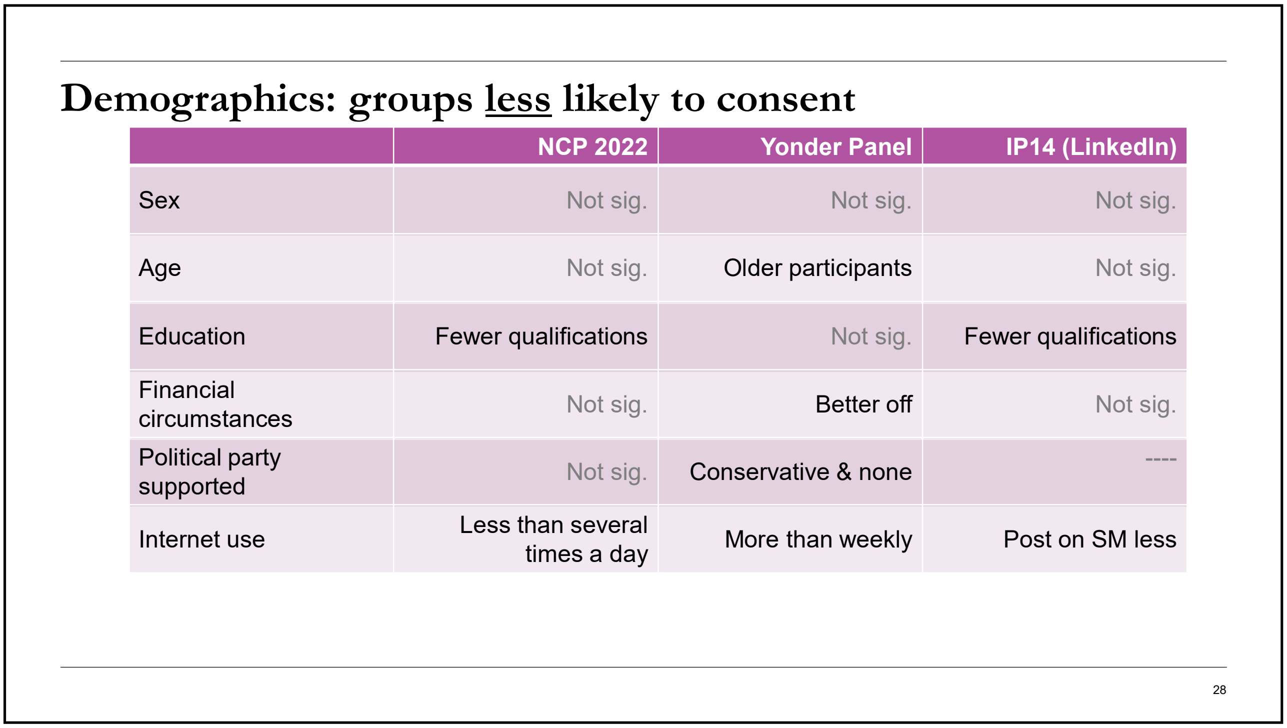
Source: Yonder Panel; Base: UK adults (18+) with a Twitter account: Not offered £2 (1,960); Offered £2 (1,968)

LinkedIn Consent

- LinkedIn - Public facing content
- Asked for consent to LinkedIn account at IP14 (2021)
- Almost all online
- 25% have LinkedIn account (n=756)
- 2 x 2 Experiment:
  - Placement – early v. late
  - Wording – additional motivational statement about importance of data or not.



27



Demographics: groups less likely to consent

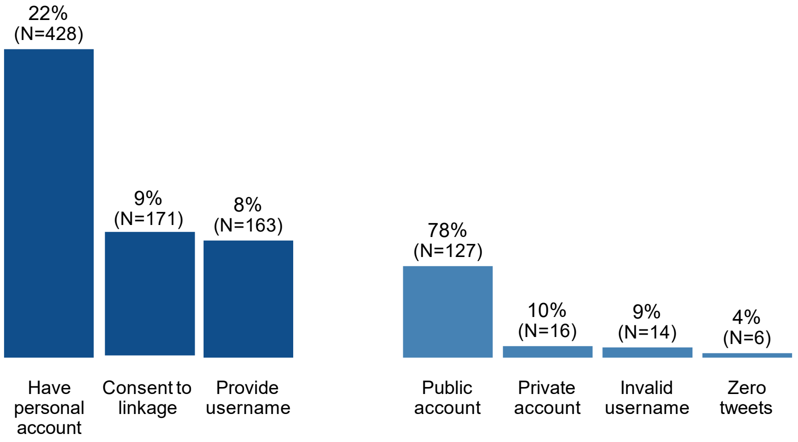
	NCP 2022	Yonder Panel	IP14 (LinkedIn)
Sex	Not sig.	Not sig.	Not sig.
Age	Not sig.	Older participants	Not sig.
Education	Fewer qualifications	Not sig.	Fewer qualifications
Financial circumstances	Not sig.	Better off	Not sig.
Political party supported	Not sig.	Conservative & none	----
Internet use	Less than several times a day	More than weekly	Post on SM less

28

4

The Nature of the Data

Respondent linkage IP10



Total Respondents: N=1,945.

Amount of Twitter Data Available

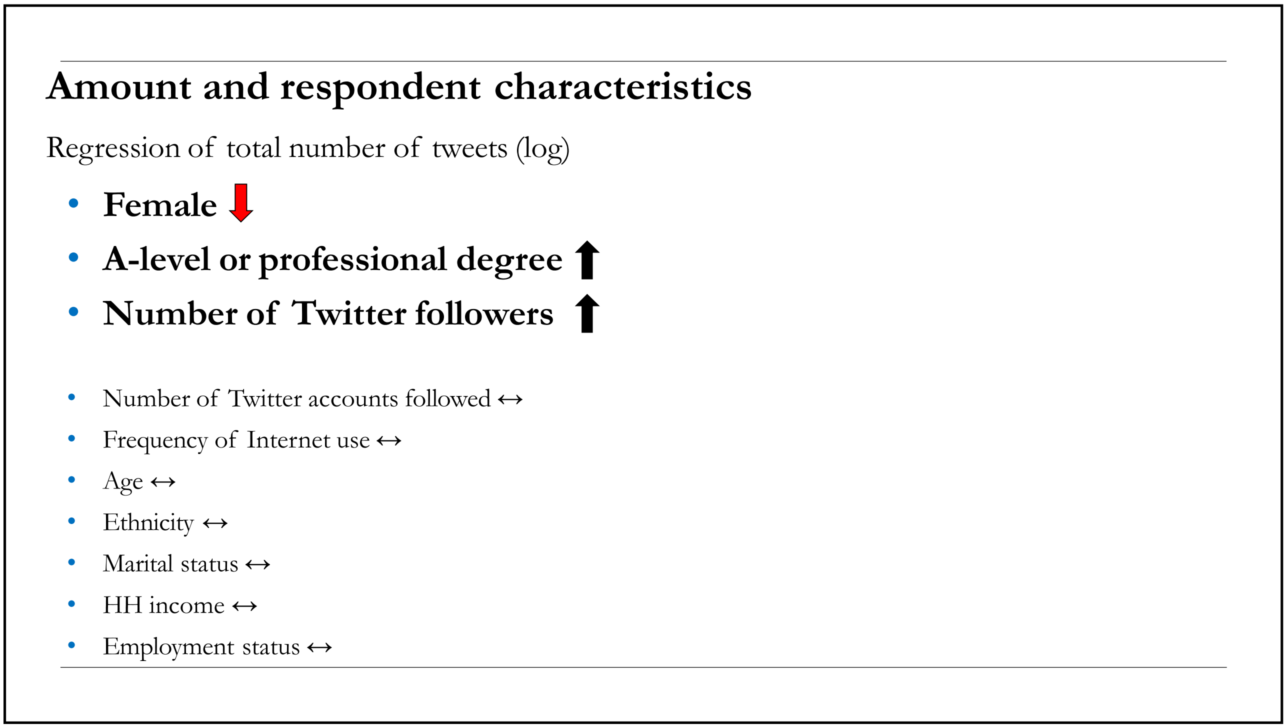
	Median	Mean	SD	Min	Max
Total tweets	**306**	**2255.32**	**6057.36**	**1**	**36451**
Followers	71	260.25	568.95	1	3734
Accounts followed	182	350.95	567.54	0	3912

31

Impact of Data Quantity

- What amount of Twitter data can be collected from respondents in a longitudinal survey?
- Amount can impact capture of signal in the noise
- Increase in variance, reduction in information
- Is there potential bias in substantive analyses?

32



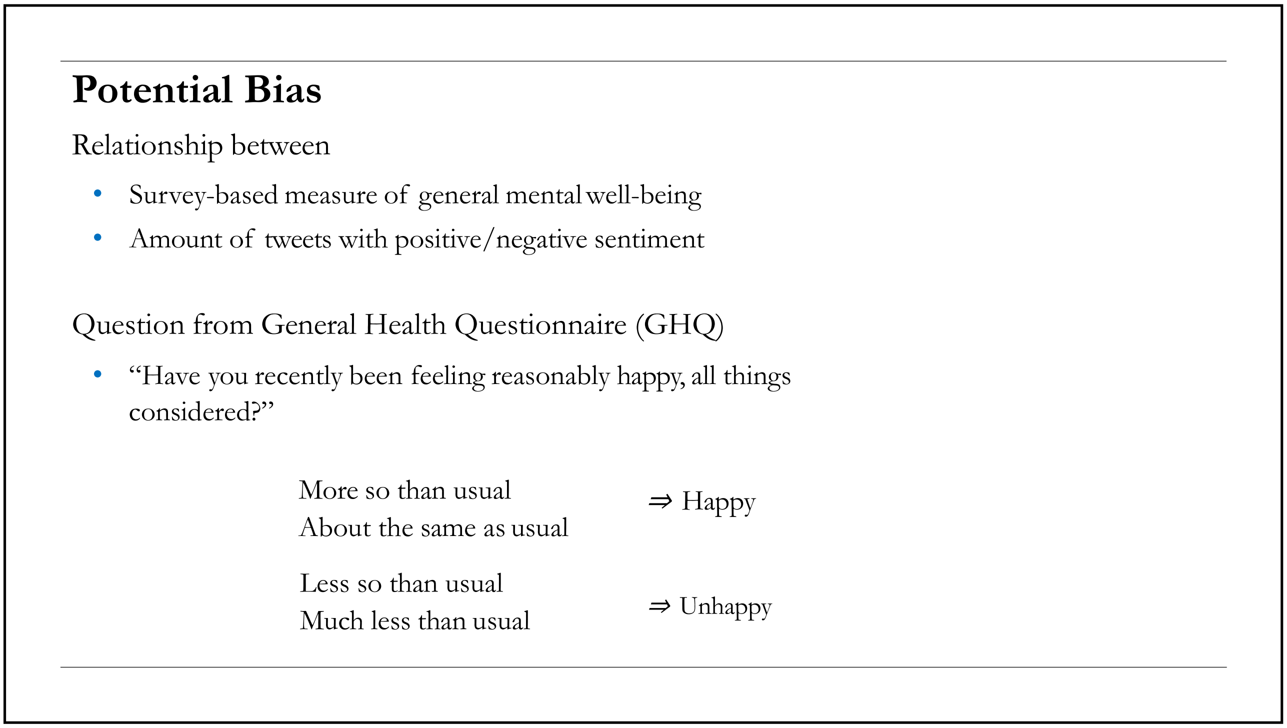
Amount and respondent characteristics

Regression of total number of tweets (log)

- **Female ⬇**
- **A-level or professional degree ⬆**
- **Number of Twitter followers ⬆**

- Number of Twitter accounts followed ↔
- Frequency of Internet use ↔
- Age ↔
- Ethnicity ↔
- Marital status ↔
- HH income ↔
- Employment status ↔

33



Potential Bias

Relationship between

- Survey-based measure of general mental well-being
- Amount of tweets with positive/negative sentiment

Question from General Health Questionnaire (GHQ)

- "Have you recently been feeling reasonably happy, all things considered?"

More so than usual About the same as usual	⇒ Happy
Less so than usual Much less than usual	⇒ Unhappy

34

Sentiment Analysis of Tweets

- Words coded +/- based on Bing lexicon (Liu 2015)
- Sentiment score calculated per tweet:
  - $sentiment = words_{positive} - words_{negative}$
- Sum of +/- tweets calculated per respondent

35

Impact of Amount on Outcomes

Regression of unhappy response

- **Number of negative tweets** ⬇
- **Number of followers** ⬆
- **Female** ⬆
- **Employed** ⬆

- Number of positive tweets ↔
- Number of accounts followed ↔
- Age ↔
- Education ↔
- Ethnicity ↔
- Marital status ↔
- HH income ↔

36

5

Archiving and Sharing

37

37

Archiving and Sharing

- Archiving and sharing of data is important:
- Replication of results
- Maximise value of data

- Particular issues:
- Who is responsible for maintaining the data?
- Deleted Tweets/withdrawn consent
  - Multiple consent requests in longitudinal survey?
- Legal issues of sharing Twitter datasets

38

Secure access to linked data

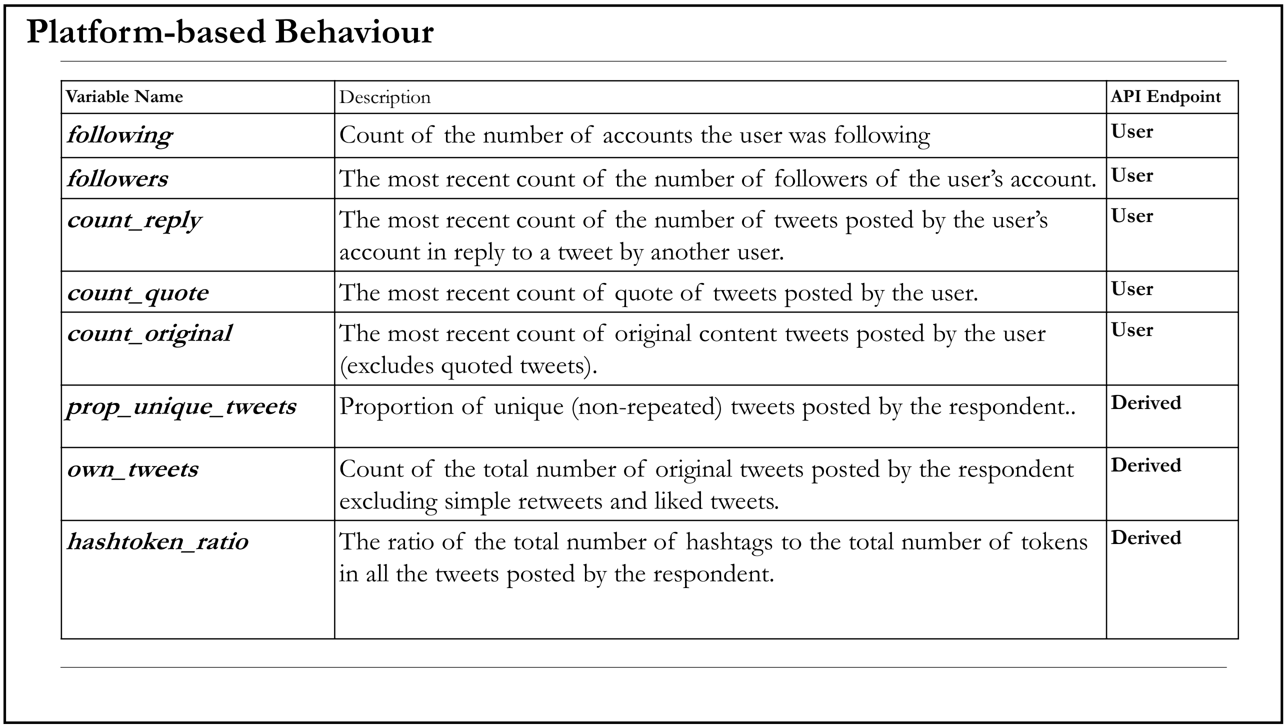
- Quasi-anonymisation & cut-down datasets
- Consideration of justification for research
- Training/accreditation of researchers
- Documentation of access
- Access to raw data in a secure environment
  - Offline access (if possible)
  - Not able to take data away (without review)

Two datasets

Platform-based behavior (raw and derived metrics from user-level metadata) [**30 variables**]

Tweet metadata (raw and derived metrics from tweet-level metadata) [**135 variables**]:

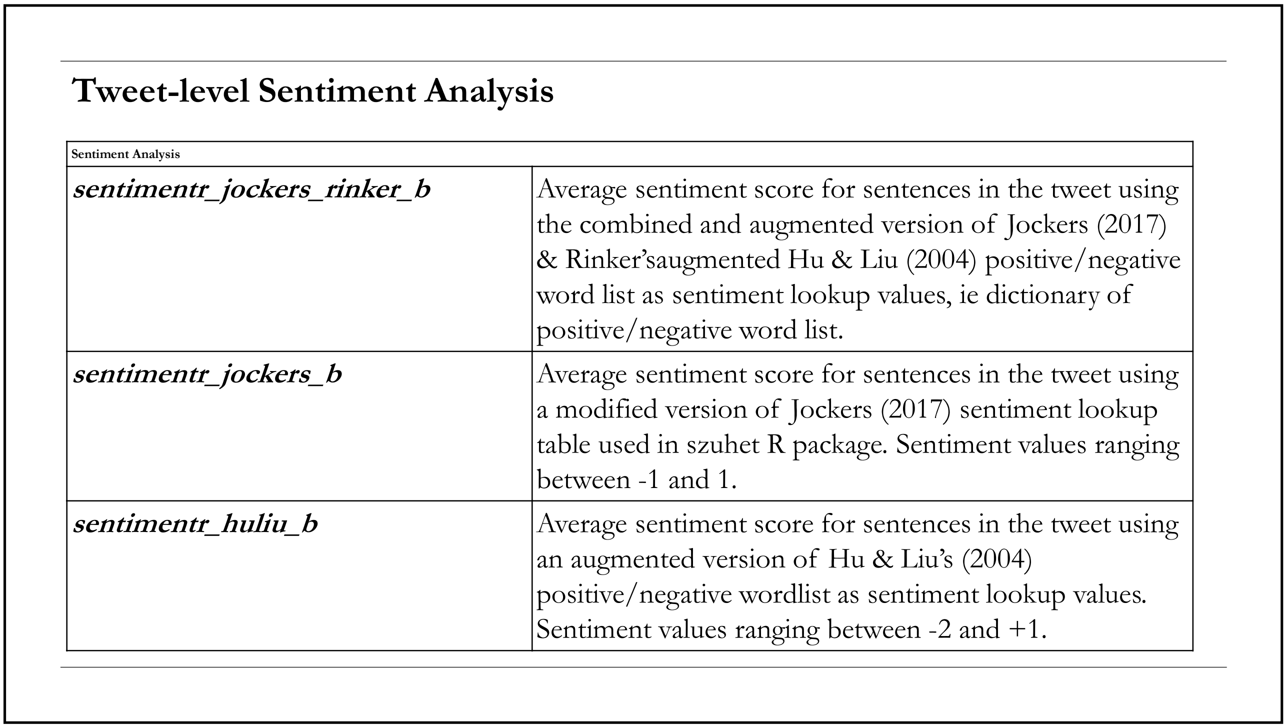
- Tweet raw metadata
- Sentiment Analysis
- Syntactic and Lexical Features
- Readability
- Lexical Diversity
- Complex content: Part-of-Speech tagging



Platform-based Behaviour

Variable Name	Description	API Endpoint
following	Count of the number of accounts the user was following	**User**
followers	The most recent count of the number of followers of the user's account.	**User**
count_reply	The most recent count of the number of tweets posted by the user's account in reply to a tweet by another user.	**User**
count_quote	The most recent count of quote of tweets posted by the user.	**User**
count_original	The most recent count of original content tweets posted by the user (excludes quoted tweets).	**User**
prop_unique_tweets	Proportion of unique (non-repeated) tweets posted by the respondent..	**Derived**
own_tweets	Count of the total number of original tweets posted by the respondent excluding simple retweets and liked tweets.	**Derived**
hashtoken_ratio	The ratio of the total number of hashtags to the total number of tokens in all the tweets posted by the respondent.	**Derived**

41



Tweet-level Sentiment Analysis

Sentiment Analysis	
sentimentr_jockers_rinker_b	Average sentiment score for sentences in the tweet using the combined and augmented version of Jockers (2017) & Rinker'saugmented Hu & Liu (2004) positive/negative word list as sentiment lookup values, ie dictionary of positive/negative word list.
sentimentr_jockers_b	Average sentiment score for sentences in the tweet using a modified version of Jockers (2017) sentiment lookup table used in szuhet R package. Sentiment values ranging between -1 and 1.
sentimentr_huliu_b	Average sentiment score for sentences in the tweet using an augmented version of Hu & Liu's (2004) positive/negative wordlist as sentiment lookup values. Sentiment values ranging between -2 and +1.

42

Tweet-level Lexical analysis

Syntactic and Lexical Features	
chars	Count of characters per tweet.
sents	Count of sentences in the tweet.
tokens	Count of tokens (words) per tweet.

Lexical Diversity	
C	Herdan's C (Herdan, 1960, as cited in Tweedie & Baayen, 1998; sometimes referred to as LogTTR)
R	Guiraud's Root TTR (Guiraud, 1954, as cited in Tweedie & Baayen, 1998)
TTR	The ordinary Type-Token Ratio

Deposit

- Reviewed by data security experts to ensure minimized risks
- Created code book on how to use
- Data processed using Understanding Society procedures
- Deposit to the UK Data Archive (soon!)
- Open access to researchers to link to the longitudinal data

5

Next Steps

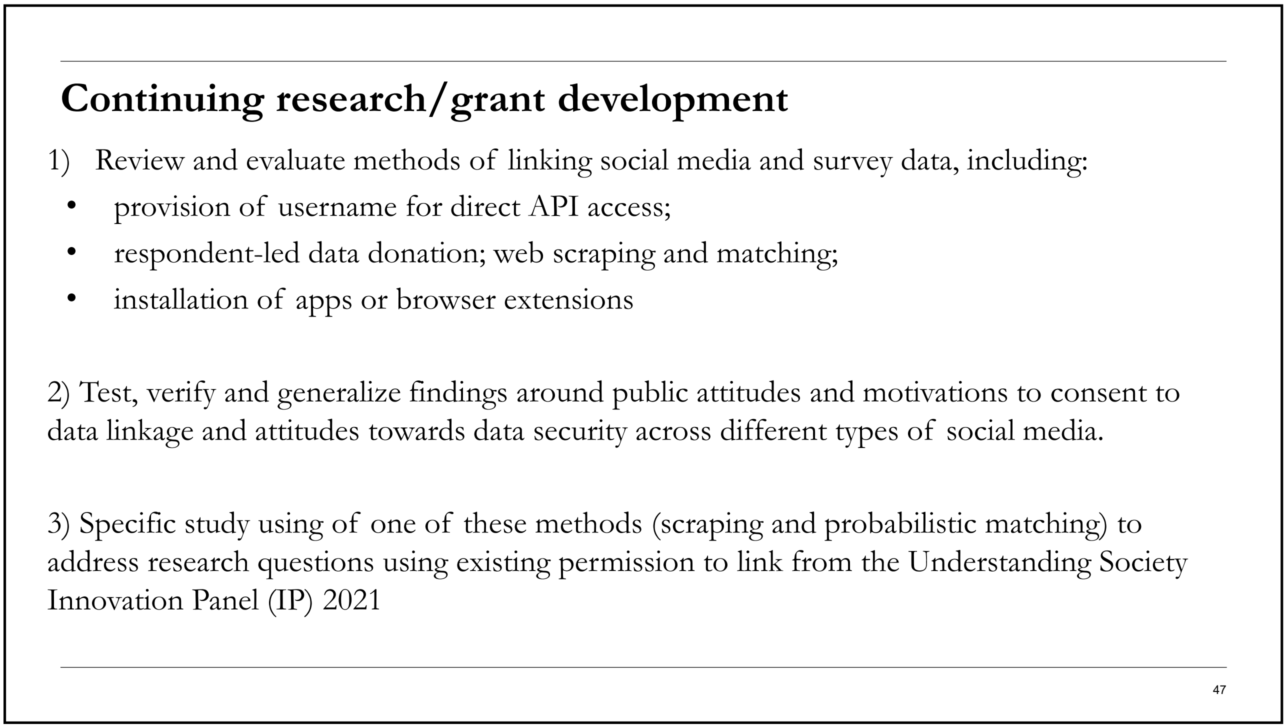
Use in Nonresponse for Longitudinal Studies

Continual, ongoing past attrition (?)

Can we use to trace?

Or use in nonresponse adjustments?

But limited to specific subgroup



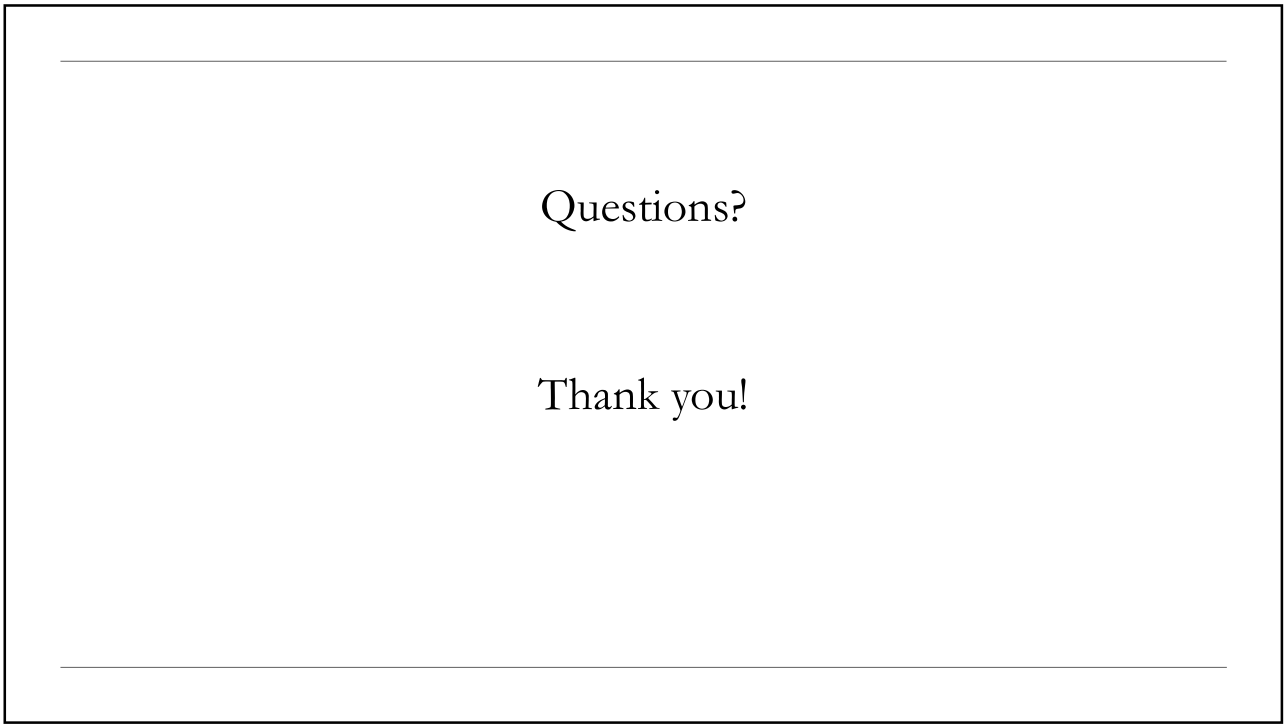
Continuing research/grant development

1) Review and evaluate methods of linking social media and survey data, including:

- provision of username for direct API access;
- respondent-led data donation; web scraping and matching;
- installation of apps or browser extensions

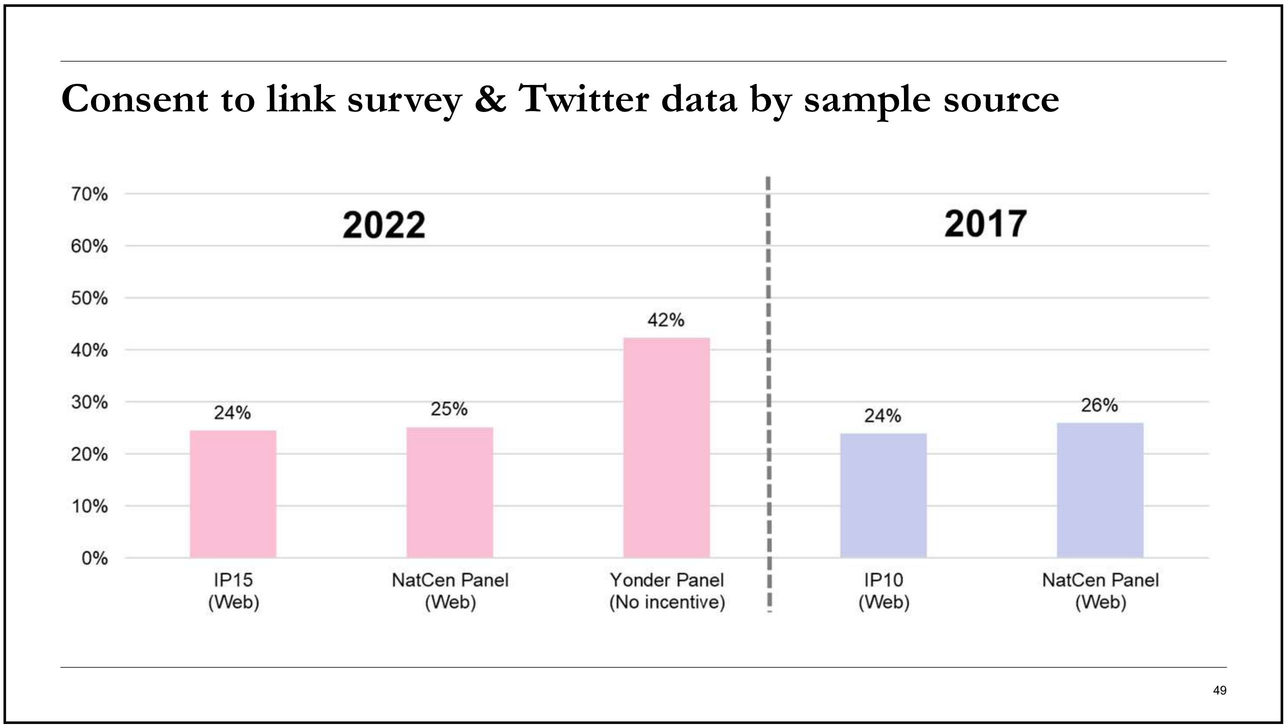
2) Test, verify and generalize findings around public attitudes and motivations to consent to data linkage and attitudes towards data security across different types of social media.

3) Specific study using of one of these methods (scraping and probabilistic matching) to address research questions using existing permission to link from the Understanding Society Innovation Panel (IP) 2021

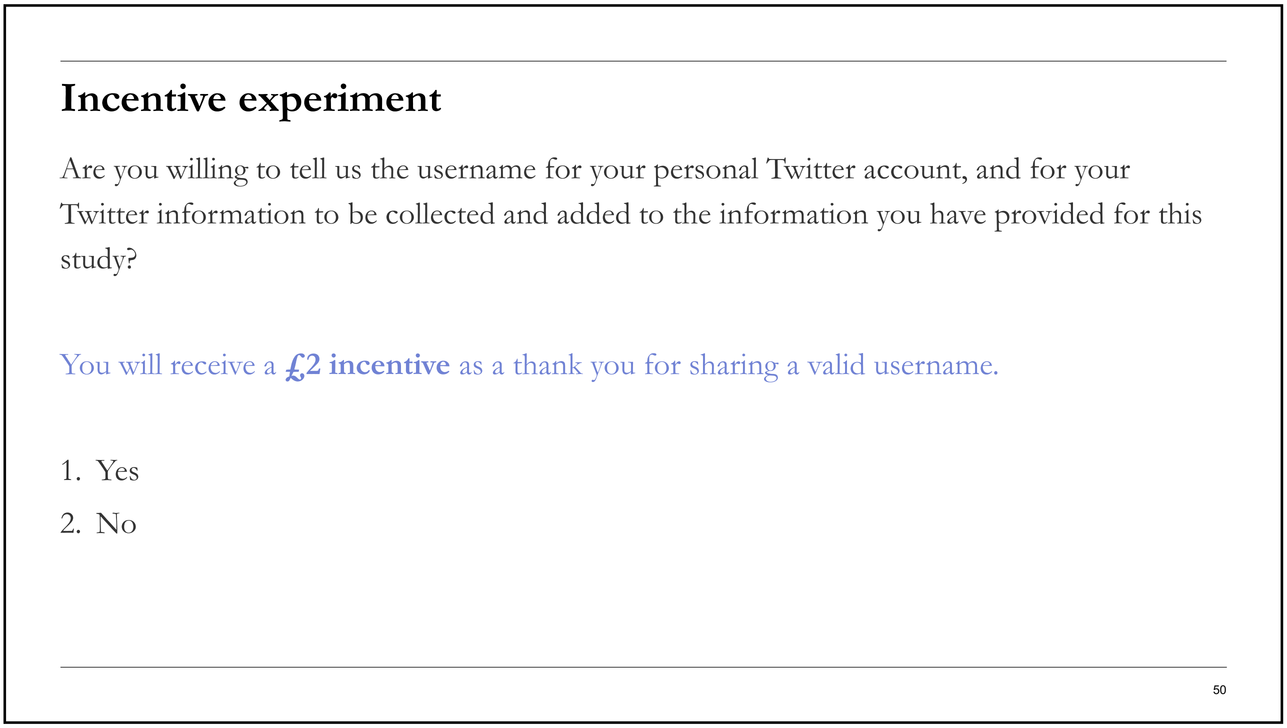


Questions?

Thank you!



Consent to link survey & Twitter data by sample source

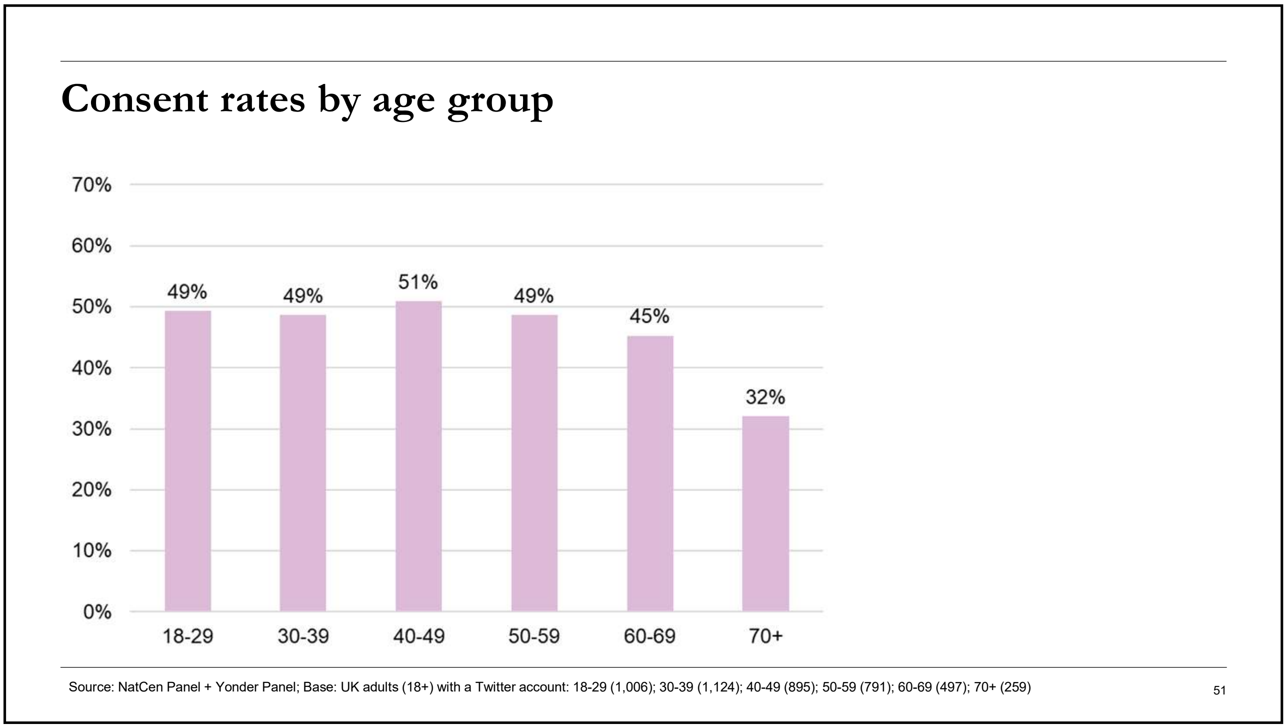


Incentive experiment

Are you willing to tell us the username for your personal Twitter account, and for your Twitter information to be collected and added to the information you have provided for this study?

You will receive a **£2 incentive** as a thank you for sharing a valid username.

1. Yes
2. No

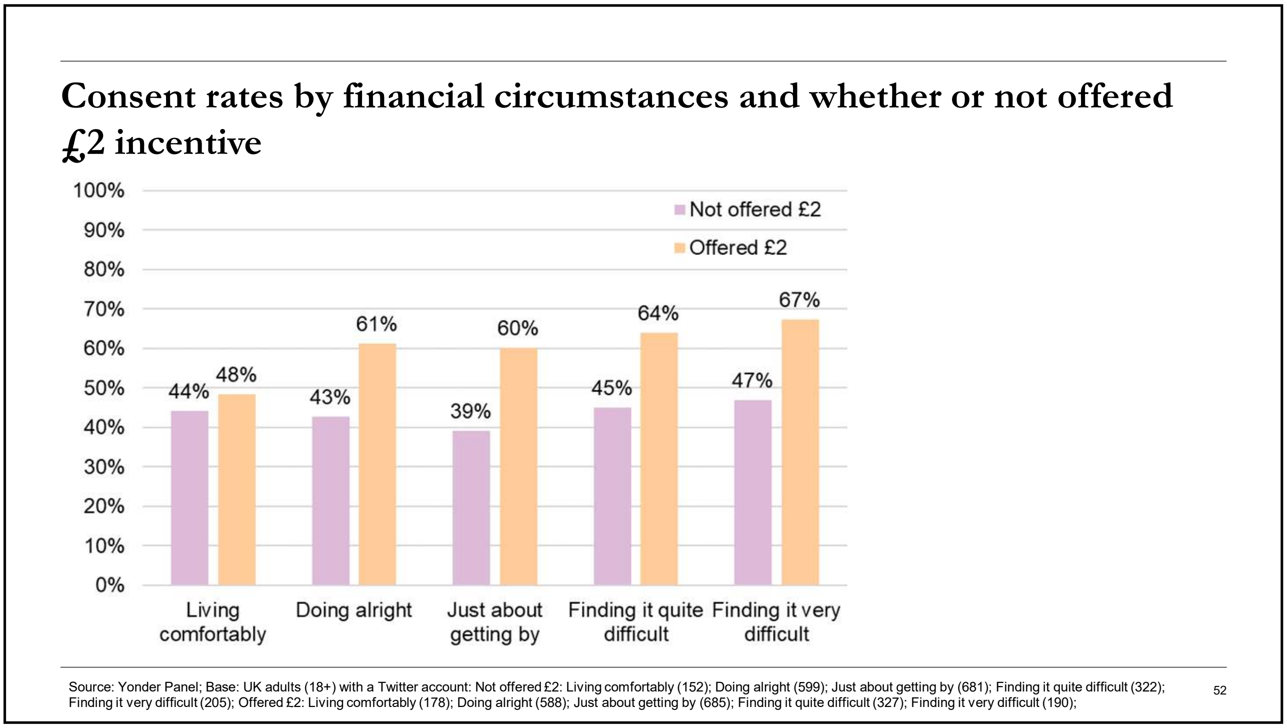


Consent rates by age group

Source: NatCen Panel + Yonder Panel; Base: UK adults (18+) with a Twitter account: 18-29 (1,006); 30-39 (1,124); 40-49 (895); 50-59 (791); 60-69 (497); 70+ (259)

51

51

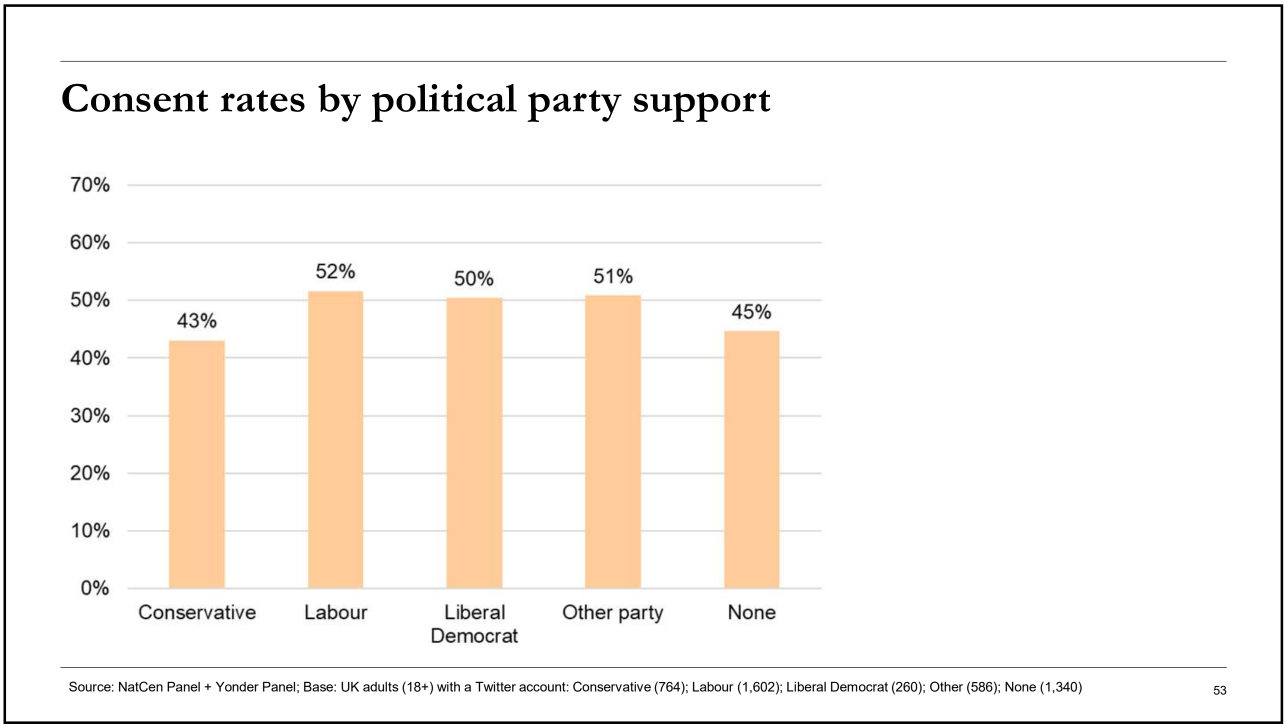


Consent rates by financial circumstances and whether or not offered £2 incentive

Source: Yonder Panel; Base: UK adults (18+) with a Twitter account: Not offered £2: Living comfortably (152); Doing alright (599); Just about getting by (681); Finding it quite difficult (322); Finding it very difficult (205); Offered £2: Living comfortably (178); Doing alright (588); Just about getting by (685); Finding it quite difficult (327); Finding it very difficult (190);

52

52

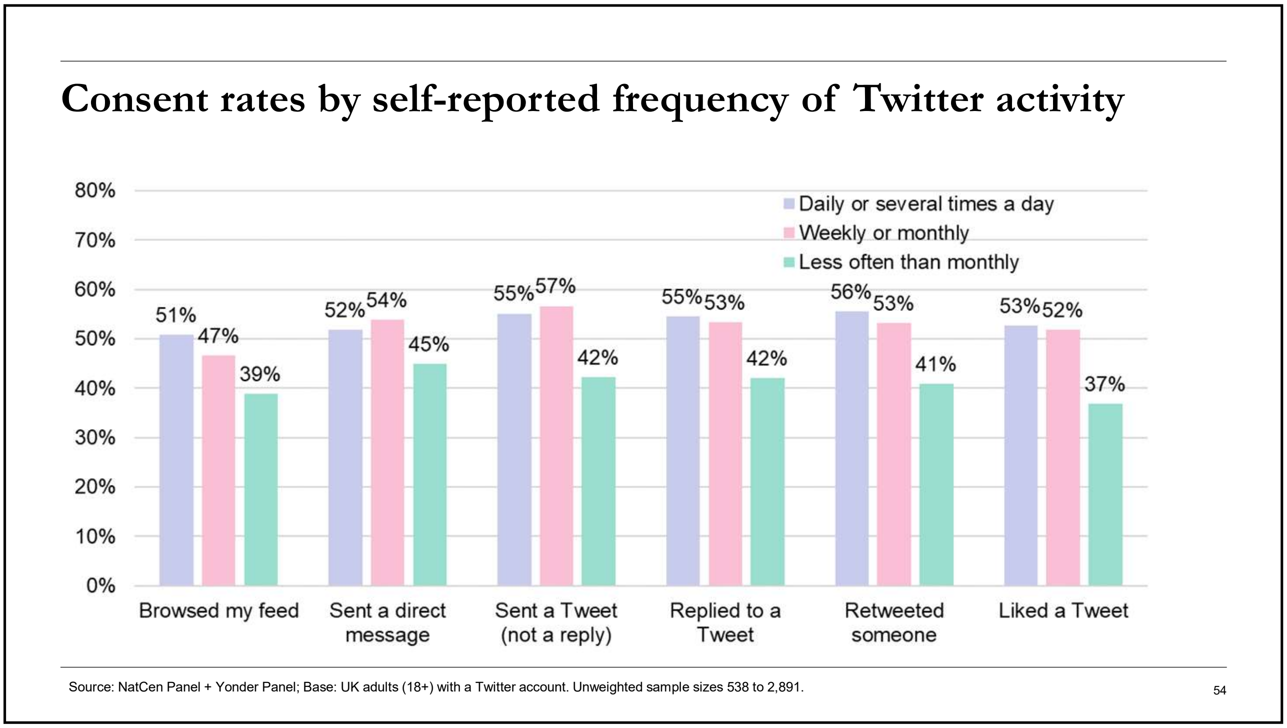


Consent rates by political party support

Source: NatCen Panel + Yonder Panel; Base: UK adults (18+) with a Twitter account: Conservative (764); Labour (1,602); Liberal Democrat (260); Other (586); None (1,340)

53

53

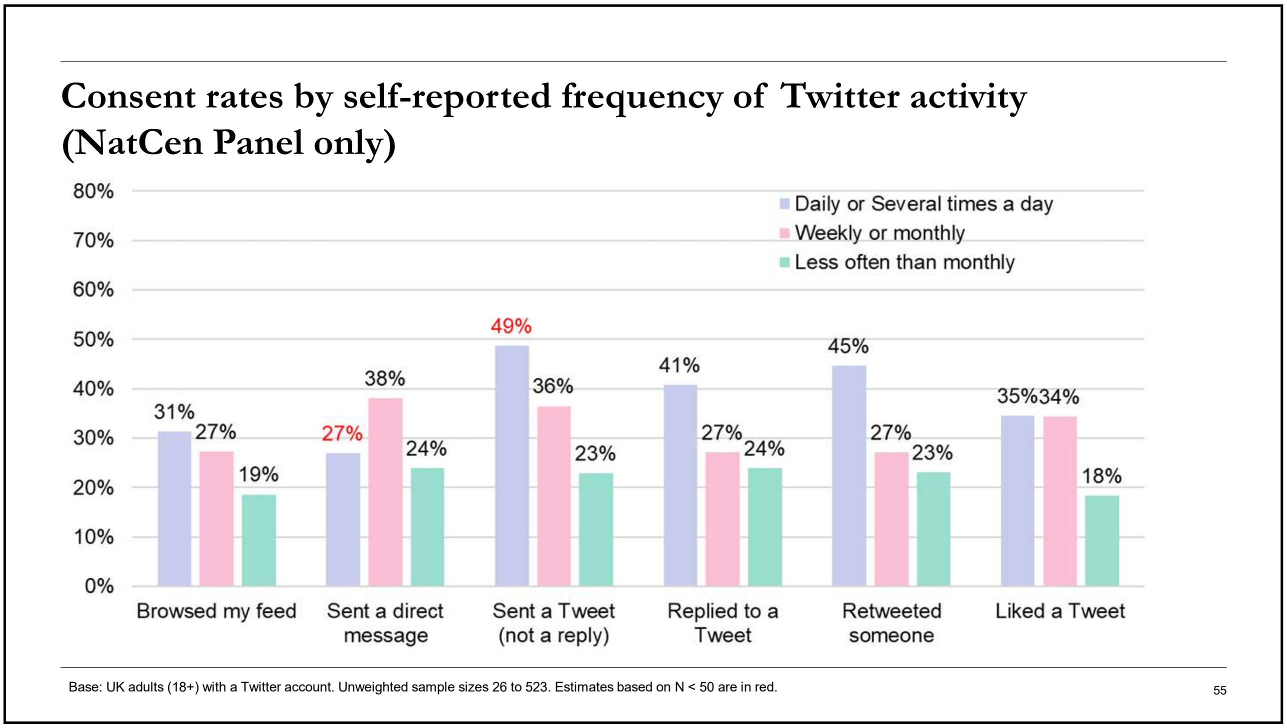


Consent rates by self-reported frequency of Twitter activity

Source: NatCen Panel + Yonder Panel; Base: UK adults (18+) with a Twitter account. Unweighted sample sizes 538 to 2,891.

54

54

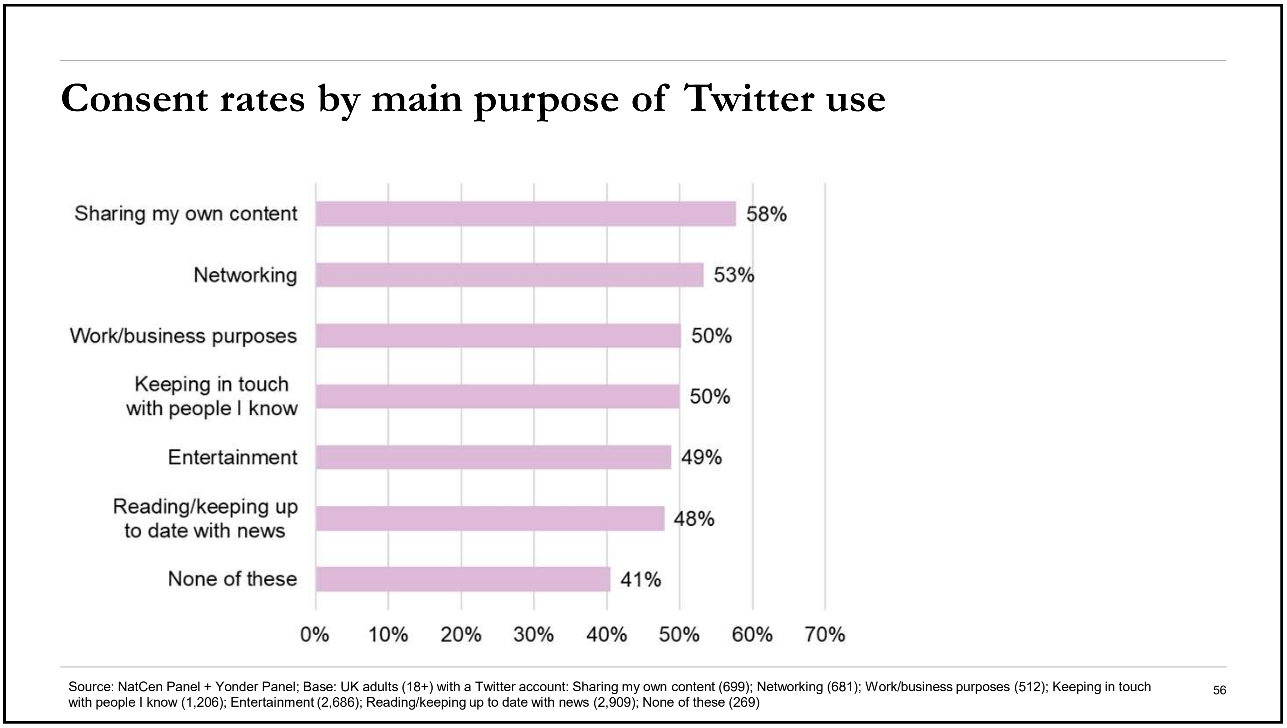


Consent rates by self-reported frequency of Twitter activity (NatCen Panel only)

Base: UK adults (18+) with a Twitter account. Unweighted sample sizes 26 to 523. Estimates based on N < 50 are in red.

55

55



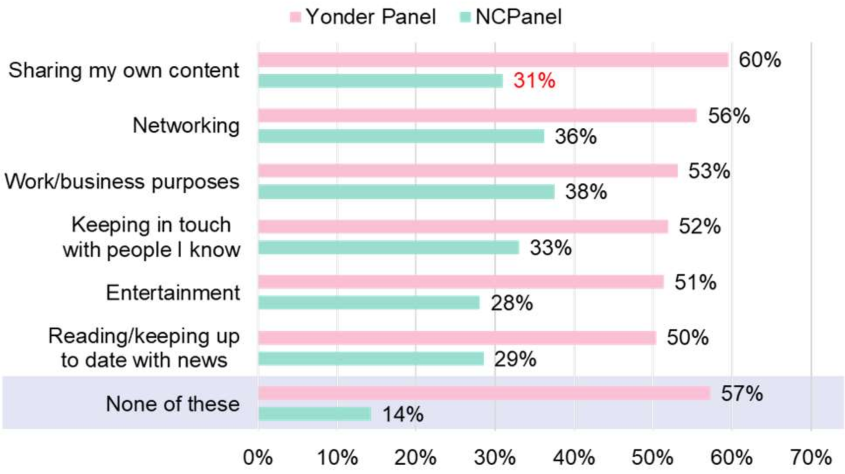
Consent rates by main purpose of Twitter use

Source: NatCen Panel + Yonder Panel; Base: UK adults (18+) with a Twitter account: Sharing my own content (699); Networking (681); Work/business purposes (512); Keeping in touch with people I know (1,206); Entertainment (2,686); Reading/keeping up to date with news (2,909); None of these (269)

56

56

Consent rates by main purpose of Twitter use and sample source



Base: UK adults (18+) with a Twitter account: Yonder unweighted sample sizes 164 to 2,577; NatCen Panel unweighted sample sizes 42 to 322. Estimates based on N < 50 are in red.

Summary & reflections (1)

- Changes to consent question wording, including positioning of additional information, does not appear to have affected consent rates
  - But the impact on how *informed* consent is is unknown.
  - Consent wording is still long, is a more dramatic change needed? Or would it continue to make no difference?

- Incentivising consent to data linkage may help improve response rates in a cost-effective manner
  - How will it work outside of non-probability web panel context?
  - Ethical considerations?

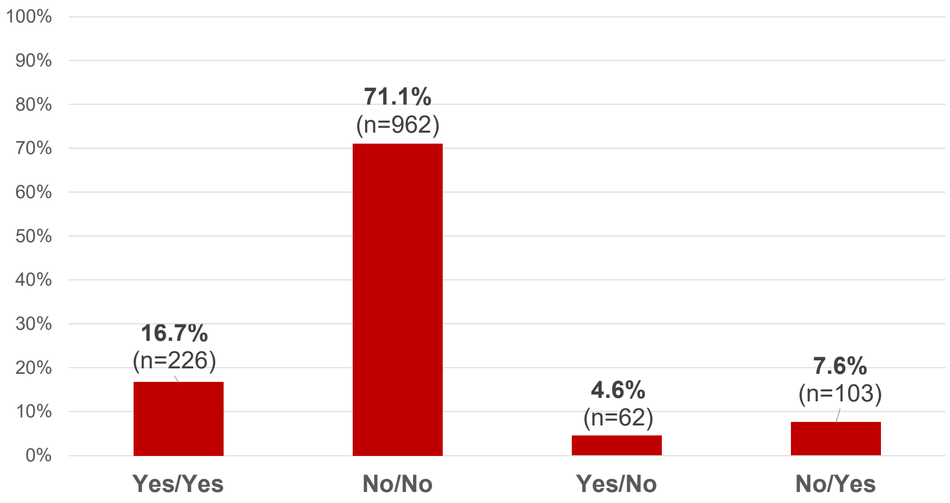
Summary & reflections (2)

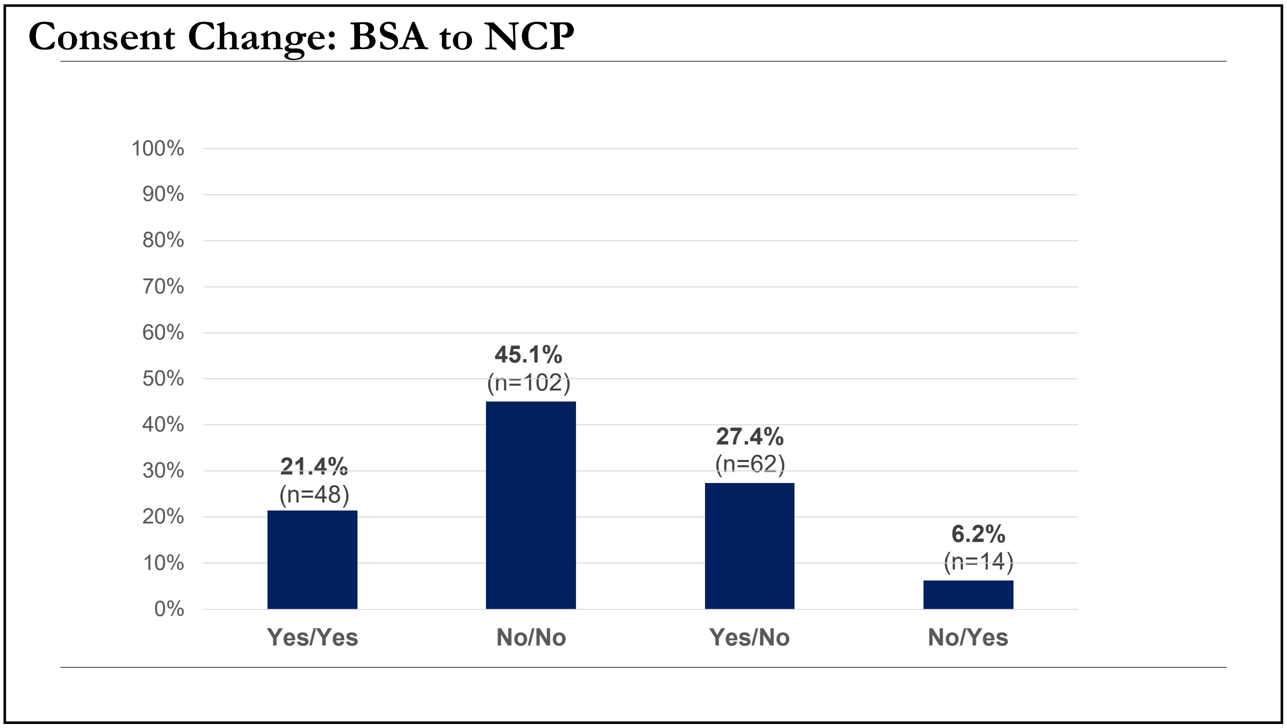
- Participants in non-probability panels appear to be more willing to consent
  - Characteristics of panel members? Nature of relationship?

- Some patterns emerging in differential consent rates:
  - Older participants, people not supporting a political party
  - In general, people who are less active on Twitter are also less likely to consent

- Moving beyond Twitter…

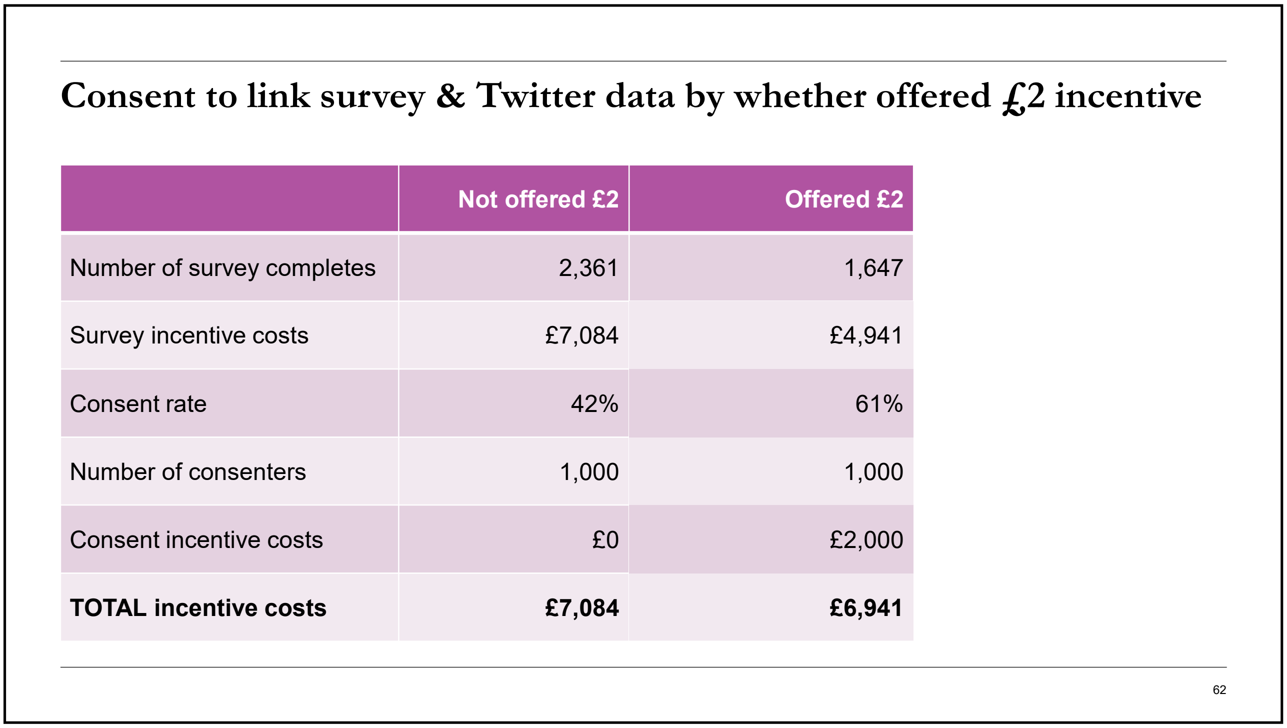
Usage Change: BSA to NCP





Consent Change: BSA to NCP

61

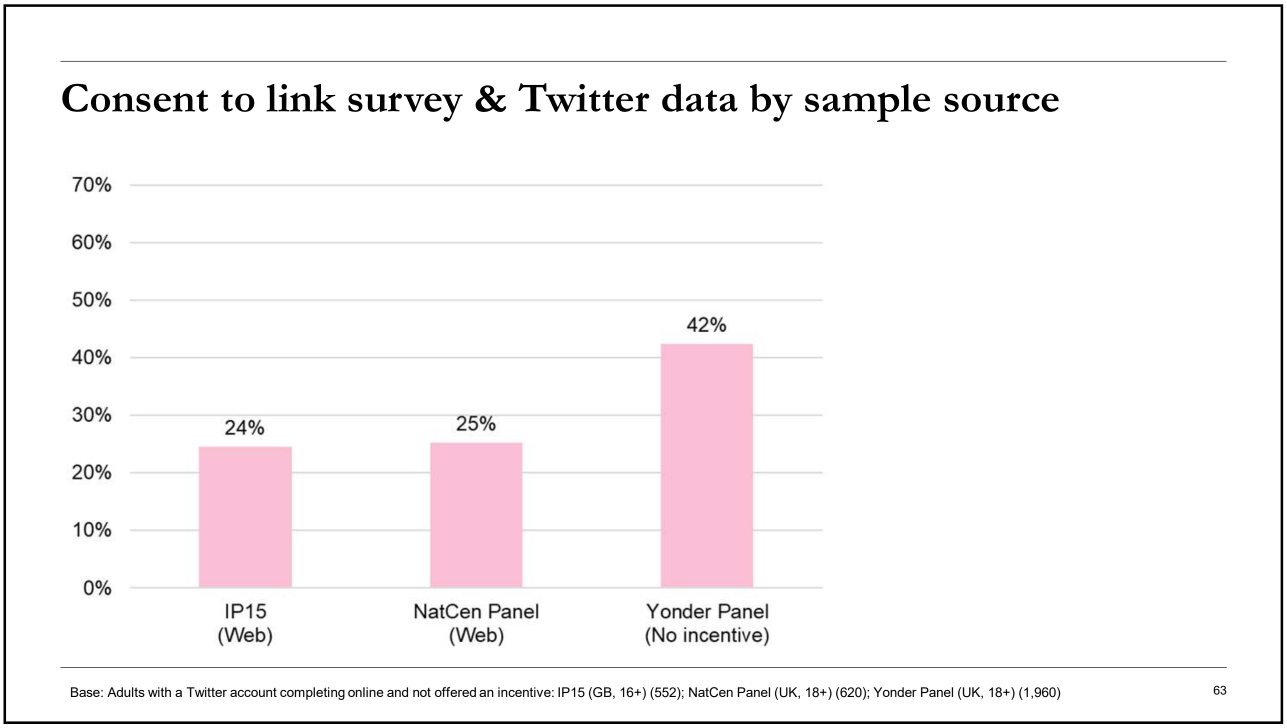


Consent to link survey & Twitter data by whether offered £2 incentive

	Not offered £2	Offered £2
Number of survey completes	2,361	1,647
Survey incentive costs	£7,084	£4,941
Consent rate	42%	61%
Number of consenters	1,000	1,000
Consent incentive costs	£0	£2,000
TOTAL incentive costs	**£7,084**	**£6,941**

62

62



Consent to link survey & Twitter data by sample source

Base: Adults with a Twitter account completing online and not offered an incentive: IP15 (GB, 16+) (552); NatCen Panel (UK, 18+) (620); Yonder Panel (UK, 18+) (1,960)

63

63