



Understanding Society

THE UK HOUSEHOLD LONGITUDINAL STUDY

Linking survey and social media data in Understanding Society

Tarek Al Baghal
ISER University of Essex



University of Essex



**Economic
and Social
Research Council**

An initiative by the Economic and Social Research Council, with scientific leadership by the Institute for Social and Economic Research, University of Essex, and survey delivery by NatCen Social Research and Kantar Public



Acknowledgments

- Curtis Jessop – NatCen Social Research
- Luke Sloan – University of Cardiff
- Alexander Wenz – University of Mannheim

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
- 79% say they use Facebook account
- 47% say they use Twitter account
 - Likely skewed – online panel data
- 33 million monthly-active FB users (2013)
- 15 million monthly-active Twitter users (2013)

What are we trying, and why?

- Link survey participants' answers to publicly available information from their Twitter accounts
- Survey data benefits from real-time, 'natural' behavioural and attitudinal data
- Between wave measures(?)

What are we trying, and why?

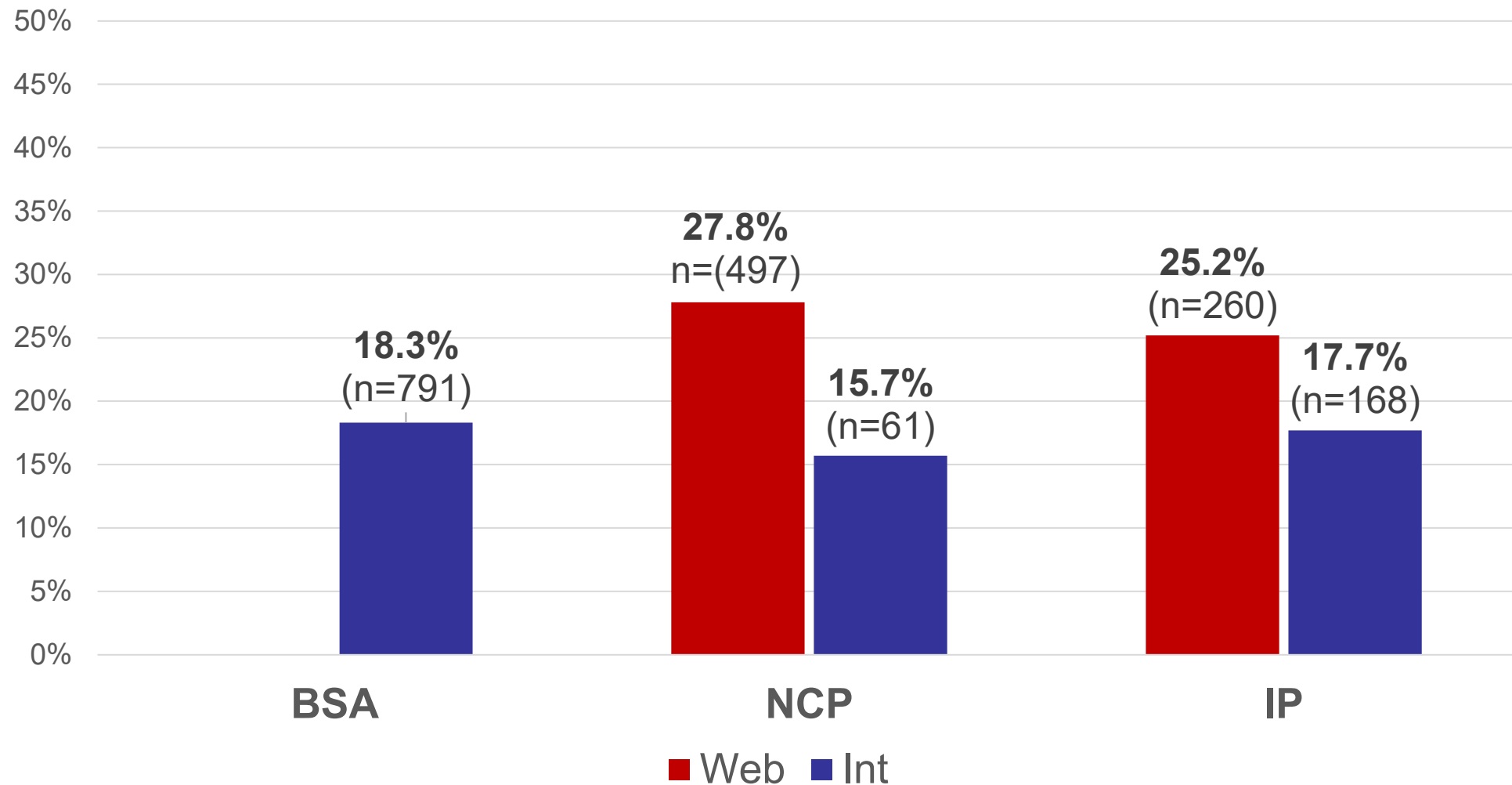
- Adds the 'who' to Twitter data – creates a sample frame
- Analysis of different groups on Twitter
- Find acceptable archiving methods
- Complement, not contrast!

Consent in 3 UK Studies

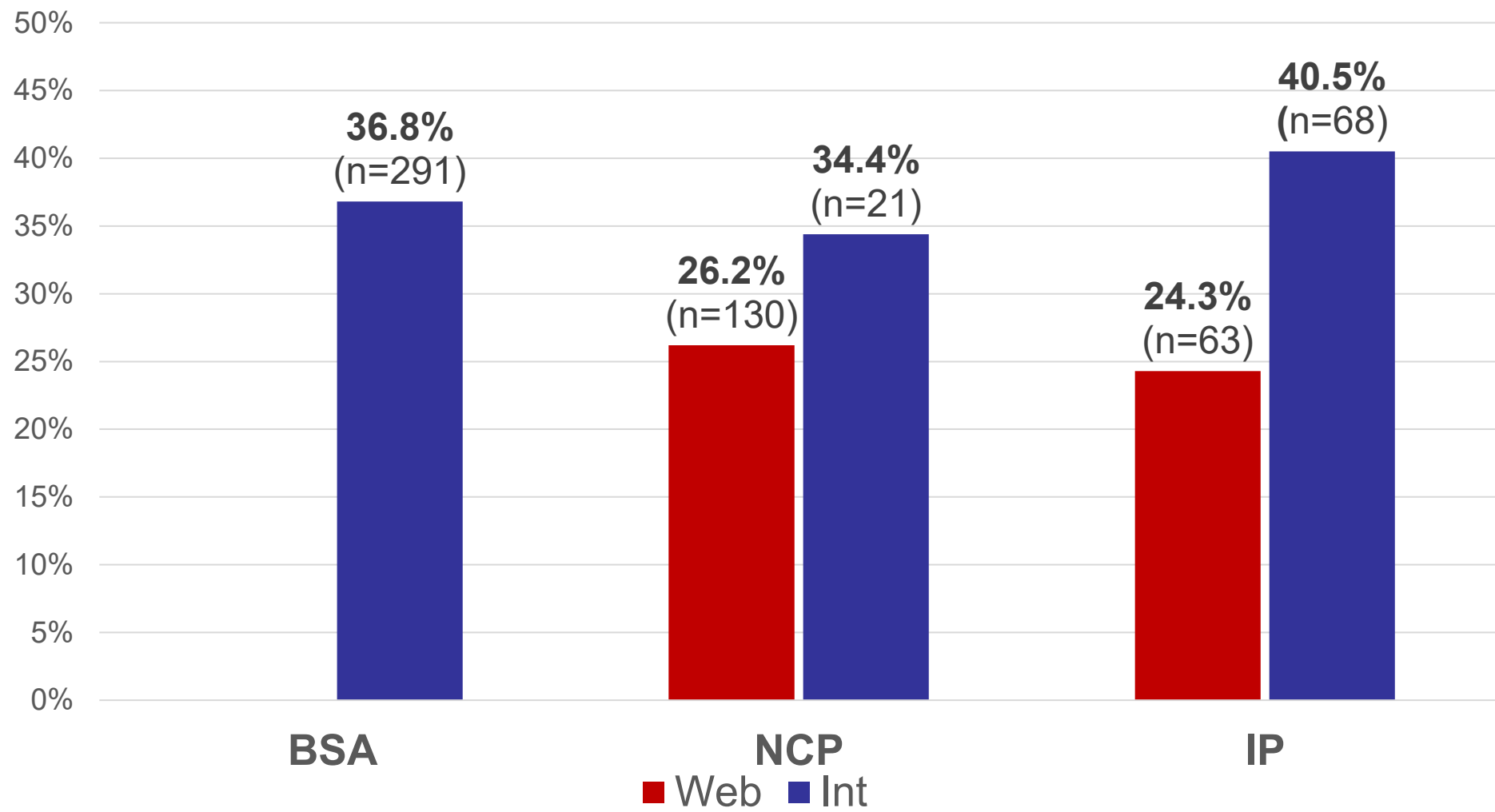
- British Social Attitudes (2015)
 - Cross-sectional, F2F
- NatCen Panel (July 2017)
 - Longitudinal, BSA 2015,2016 samples
 - Sequential MM Web - CATI
- Innovation Panel (2017)
 - Longitudinal, annual since 2008
 - CAPI, & Sequential MM Web - CAPI



How many Twitter Users?



Consent Among Users



Significant Characteristics

- British Social Attitudes (2015)
 - Age: Younger ↑ Older ↓
- NatCen Panel (July 2017)
 - Age: Younger ↑ Older ↓
 - Sex: Male ↑ Female ↓
- Innovation Panel (2017)
 - Mode: F2F ↑ Web ↓


ISER
 @iseressex · FOLLOWS YOU

TWEETS 2,028
 FOLLOWING 1,065
 FOLLOWERS 2,310
 FAVOURITES 92
 LISTS 2


Following

@iseressex · FOLLOWS YOU

World leaders in the production and analysis of longitudinal surveys & microsimulation. Home to MiSoC @usociety & @EUROMOD_

University of Essex, UK
iser.essex.ac.uk
 Joined June 2009

Tweet to Message

80 Followers you know







38 Photos and videos








ISER retweeted

UnderstandingSociety @usociety · Jul 7

Take our grip strength test at #esra15 understandingsociety.ac.uk/2015/07/07/sur... @ESRASurvey @iseressex @DIW_Berlin_en @uni_iceland



View photo

ISER @iseressex · Jul 7

How reforms come about - a presentation on recent changes to family support in Poland: using @EUROMOD_ evidence

ISER @iseressex · Jul 7

Day2 of our @EUROMOD_ meet in Athens: how the Austrian team worked with their finance ministry on tax reform using our @EUROMOD_ model ...

Survey Magazine @SurveyMagazine · Follow

Discover Society @DiscoverSociety · Follow

CPC @CPCpopulation · Follow

Find friends

Mexico City Trends · [Change](#)

#SabadoDeGanarSeguidores
 #BatmanvSuperman
 #EsteFinde
 Green Lantern Corps
 #5HReflectionForNewSingle
 Serena Williams
 #MiCorazonEsUnPais
 Matteo Darmian
 Funes Mori
 #T02015

© 2015 Twitter · [About](#) · [Help](#) · [Ads info](#)

Source

Console

```
> timeline[1:8]
[[1]]
[1] "iseressex: RT @usociety: Take our grip strength test at #esra15 https://t.co/6AcElcEjjo @ESRASurvey @iseressex @DIW_Berlin_en @uni_iceland http://t.co..."

[[2]]
[1] "iseressex: How reforms come about - a presentation on recent changes to family support in Poland: using @EUROMOD_ evidence"

[[3]]
[1] "iseressex: Day2 of our @EUROMOD_ meet in Athens: how the Austrian team worked with their finance ministry on tax reform using our @EUROMOD_ model ..."

[[4]]
[1] "iseressex: RT @cernat_a: @iseressex at European Survey Research Association Conference #ESRA15, Reykjavik, 13-17 July https://t.co/FKzEfV6aTD"

[[5]]
[1] "iseressex: Our new Policy Effects tool as part of @EUROMOD_ makes checking policies for their plausibility much quicker: Prof Holly Sutherland wraps up"

[[6]]
[1] "iseressex: An amazing day for a conf in Athens - & resrch on Greek (& other countries) policies v topical & relevant - grt work by @EUROMOD_ analysts"

[[7]]
[1] "iseressex: Chrysa Leventi presents her evaluation of guaranteed min income policy in Greece to protect against poverty http://t.co/XniYMQ98C @IMFNews"

[[8]]
[1] "iseressex: But true picture of impact of cumulative policy changes on #Italy people only just starting to emerge - good inter generation redistribution"

> followers[1:5]
$`3303562877`
[1] "aysegulkayaoglu"

$`39972284`
[1] "WeClearJunk"

$`20409361`
[1] "Srathod"

$`14340145`
[1] "usabilitycounts"

$`70479696`
[1] "PeanutLabsMedia"
```

Environment

History

Import Dataset

Global Environment

```
Environment
iseressex
description : chr "World leaders in the production and analysis of longitudinal surveys & mi..."
statusesCount : num 2028
followersCount : num 2310
favoritesCount : num 92
friendsCount : num 1065
url : chr "http://t.co/12K6TtCciv"
name : chr "ISER"
created : POSIXct[1:1], format: "2009-06-19 11:08:50"
protected : logi FALSE
verified : logi FALSE
screenName : chr "iseressex"
location : chr "University of Essex, UK"
lang : chr "en"
id : chr "48676963"
lastStatus : Reference class 'status' [package "twitter"] with 17 fields
..$ text : chr "RT @usociety: Take our grip strength test at #esra15 https://t.co/6AcElcEjjo..."
..$ favorited : logi FALSE
..$ favoriteCount: num 0
..$ replyToSN : chr(0)
..$ created : POSIXct[1:1], format: "2015-07-08 10:55:24"
..$ truncated : logi FALSE
..$ replyToSID : chr(0)
..$ id : chr "618735125128458240"
..$ replyToUID : chr(0)
..$ statusSource : chr "<a href='\"http://twitter.com/\"' rel='\"nofollow\"'>Twitter Web Client</..."
..$ screenName : chr "Unknown"
..$ retweetCount : num 2
..$ isRetweet : logi TRUE
..$ retweeted : logi FALSE
..$ longitude : chr(0)
..$ latitude : chr(0)
..$ urls : 'data.frame': 1 obs. of 5 variables:
.. ..$ url : chr "https://t.co/6AcElcEjjo"
.. ..$ expanded_url : chr "https://www.understandingsociety.ac.uk/2015/07/07/survey-team"
.. ..$ display_url : chr "understandingsociety.ac.uk/2015/07/07/sur" | __truncated__
.. ..$ start_index : num 53
.. ..$ stop_index : num 76
.. and 51 methods of which 39 are possibly relevant.
```

Files Plots Packages Help Viewer

Programs for Coding the Data

- Tidyverse, tm (text mining), sentiment in R
- Linguistic Inquiry and Word Count (LIWC)
 - Extensive built-in dictionary
 - Text mines, outputting 80 variables.
 - Linguistic
 - Psychological
 - Social and biological processes
 - Beliefs
 - Socio-economic issues



Comparing Data Sources

- Can compare at micro-level
- Measures may show similarity, dissimilarity

E.g, partner status

- Are surveys gold standard in this case?
 - Who is discordant?
-

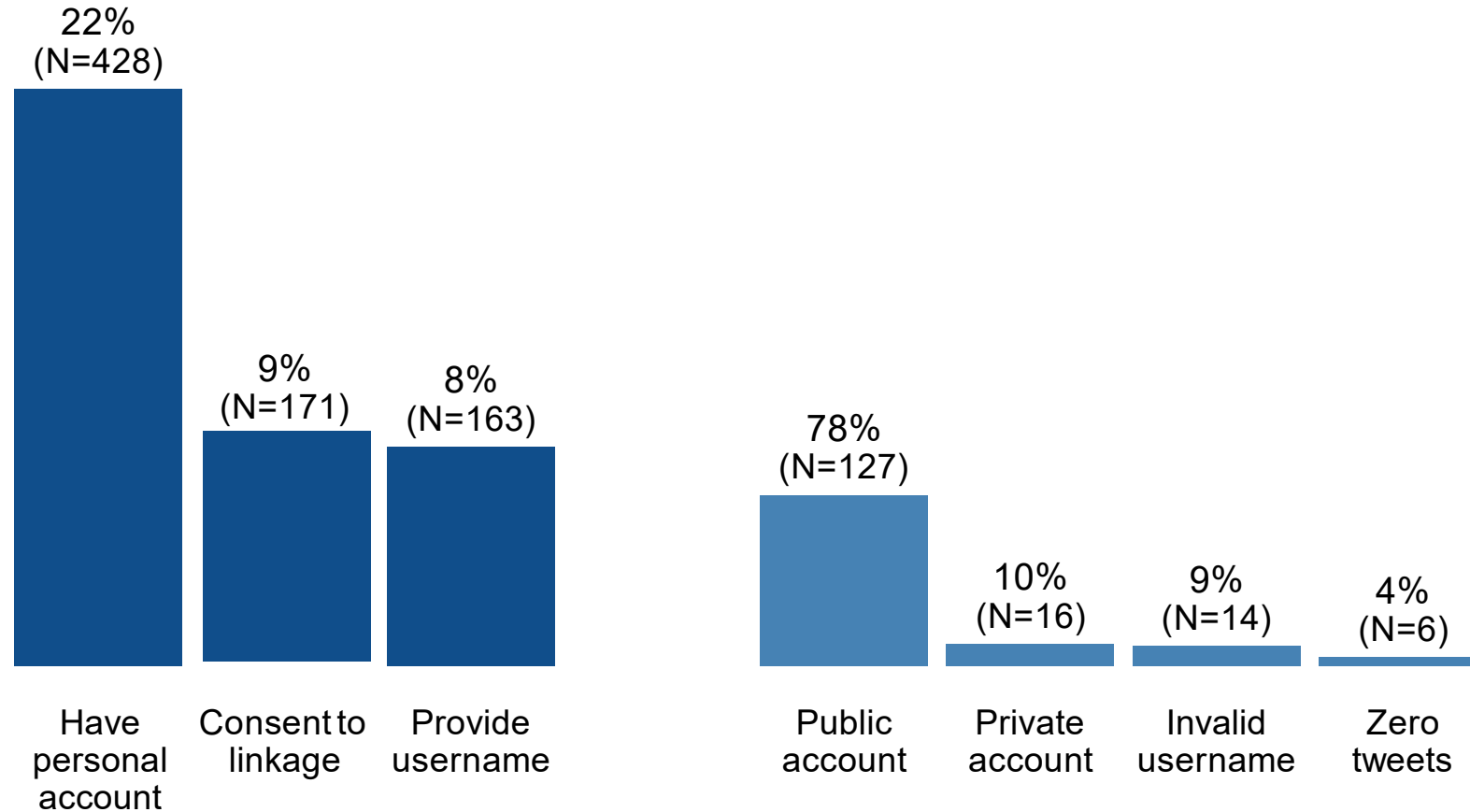
New measures

- Can create new measures, e.g.
 - Social networks, strength of ties
 - Attitudes, “likes”
 - Change on more continuous dimension
 - Reduction of respondent burden?
 - Relate to captured survey measures
-

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?

Respondent linkage IP



Total Respondents: N=1,945.

Amount of Twitter Data Available



	Median	Mean	SD	Min	Max
Scraped tweets	304	933.63	1157.60	1	3199
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

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 ↔

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

Less so than usual

Much less than usual

⇒ Happy

⇒ Unhappy

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
-

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 ↔
-

More practical issues

- Twitter allows for limited time-frames in aggregate
- Limited number of tweets.
 - Last 3200 from user, 5000 w/ a given keyword
- No control on content
- The obsolescence of Twitter (or others)?

Secure access to linked data

Existing process to access to identifiable survey 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)

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
-

Future Plans

- Long-term: Archive, Expand to main USoc
- Improve Survey Measures
 - New Measures and Error
 - Item/Unit Nonresponse
- Improve Twitter Measures
 - E.g. Improve Prediction
- Use Knowledge for Expanded Study
 - Focus on Consent
 - And target users, specific topic survey

