Received: 14 March 2021 | Accepted: 23 July 2022

DOI: 10.1111/rssa.12949

ORIGINAL ARTICLE

ROYAL STATISTICAL SOCIETY DATA | EVIDENCE | DECISIONS
Series A Statistics in Society

Evaluation of consent to link Twitter data to survey data

Zeina Mneimneh

Survey Research Center, Institute for Social Research, University of Michigan (previously), Ann Arbor, Michigan, USA

Correspondence
Zeina Mneimneh, Survey Research Center, Institute for Social Research, University of Michigan (previously), Ann Arbor, MI, USA.
Email: zeinam@umich.edu

Funding information
Michigan Institute for Data Science and Data Acquisition for Data Science, University of Michigan; National Science Foundation, Division of Social and Economic Sciences, Grant/Award Number: 1259985

Abstract

This study presents an initial framework describing factors that could affect respondents' decisions to link their survey data with their public Twitter data. It also investigates two types of factors, those related to the individual and to the design of the consent request. Individual-level factors include respondents' attitudes towards helpful behaviours, privacy concerns and social media engagement patterns. The design factor focuses on the position of the consent request within the interview. These investigations were conducted using data that was collected from a web survey on a sample of Twitter users selected from an adult online probability panel in the United States. The sample was randomly divided into two groups, those who received the consent to link request at the beginning of the survey, and others who received the request towards the end of the survey. Privacy concerns, measures of social media engagement and consent request placement were all found to be related to consent to link. The findings have important implications for designing future studies that aim at linking social media data with survey data.

KEYWORDS

consent, link, placement, privacy, social media, Twitter

1 | INTRODUCTION

Social media provides a massive amount of information on the everyday activities, opinions, thoughts, emotions and behaviours of individuals in near real-time. For example, every day, around 500 million tweets are produced (Internet Live Stats, 2021) on a wide range of topics including well-being, politics, employment, sports and leisure activities. The amount and diversity of information shared on social media constitute a source of data that could be extremely valuable to social science research. However, the scientific utility of these data requires careful attention to their unique features and investigation of their properties.

Linking social media data to another source of data with properties that can be measured or controlled (such as administrative data or survey data) can inform the potential utility of social media data for social science research. The essential premise is that information available from data source A, such as survey data, could be leveraged to investigate social media data properties. Data linkage, especially when done at an individual level—meaning an individual's data from source A is directly or indirectly linked with that same individual's data from social media, such as Twitter posts—requires obtaining the respondent's consent. While there continues to be a for/against debate about whether informed consent needs to be collected from social media users when their public data are used for research purposes (Benton et al., 2017; Fiesler & Proferes, 2018; Lane et al., 2014; Sloan et al., 2020; Zimmer, 2010), we believe that when social media data are linked to another source of data, informed consent is needed (Mneimneh, 2021; Mneimneh et al., 2020). While much remains to be discussed and debated about informed consent when using social media data, in this paper we mainly focus on understanding predictors of consent to link for the ultimate objective of maximising the value of linked data while adhering to the proper ethical conduct of social media research.

Research investigating individual-level consent to link social media data with other data sources has recently started to emerge. The majority of this research describes consent to link rates and only a few investigate a limited set of potential correlates of consent. Generally, consent to link rates are less than 50%, with the exception of Wojcik and Hughes (2019), who reported a consent to link rate of 90% among all active Twitter users of a nationally representative online panel in the United States. When consent to link is collected using a web mode, rates are generally in the range of 23%–27% (Al Baghal et al., 2019; Henderson et al., 2019; Mneimneh et al., 2020). Higher rates of up to 45% are reported for interviewer-administered modes (Al Baghal et al., 2019; Mneimneh et al., 2020). In addition to the mode of consent administration, variation in consent to link rates could depend on whether the respondents are part of a panel or not. Once consented, respondents are typically asked to log into their social media account or to report their username or handle. When self-reported, not all usernames or handles are useable as some respondents report non-valid information. Hughes et al. (2020) reported a 14 percentage point decline in consent rate when non-valid usernames were removed. This further reduces the proportion of respondents whose social media data could be linked to their survey data. Thus, investigating correlates of consent, especially those that are modifiable or that could be used for stratifying the sample to maximise consent, is extremely valuable for uncovering the utility of social media data, and for such data to live up to their full potential for social science research.

At the time of writing this paper, there is no established framework on factors that affect consent to link social media data to other data sources such as survey data. Most existing research has focused mainly on respondents' socio-demographics and the mode of consent request (Al Baghal et al., 2019; Henderson et al., 2019). Future research studies that rely on such linkage would

benefit from understanding the effect of a wider range of factors. In this paper we take a first attempt at proposing three sets of factors that could affect the decision to consent to link social media data with survey data by borrowing from factors identified in the administrative data linkage literature and others investigated in social media linkage studies.

The first set of factors is related to the individual. Individual-related factors such as altruistic traits, privacy concerns and the relevance of the source of data to the user have been theorised to be correlated with the decision to link survey data with administrative data such as medical records or social security records (Beninger et al., 2017; Jenkins et al., 2006; Sakshaug et al., 2012; Sala et al., 2012, 2014). These factors could also be related to respondents' decisions to consent to link survey data with their public social media data. Mneimneh et al. (2020) found a positive correlation between a respondent's engagement with helpful behaviours (such as offering one's seat to someone on a bus or public place, carrying a stranger's belongings and lending money to another person) and consenting to link survey data with public Twitter data in a web-administered survey among a general US population sample. The effect, however, was small and marginally significant.

Respondent's privacy concerns have been operationalised through different indirect indicators such as refusal to answer sensitive items such as income and alcohol use (Jenkins et al., 2006; Mneimneh et al., 2020; Sala et al., 2012), resistance to participate in earlier waves of a panel study, or interviewer observations of the respondent's uncooperative behaviour (Sakshaug et al., 2012). These proxy measures of privacy have been found to be negatively correlated with consent to link survey data with administrative data or Twitter public data. However, neither a direct general measure of privacy concern nor one that is specific to the type of data source has been investigated before.

Relevance of the linked data source (i.e. medical records, social security, Twitter) to the respondent has been also measured in different ways and investigated in several data linkage studies. For example, when consenting individuals to link their survey data to their medical records, relevance of the linking request has been measured by asking respondents about their health status (Woolf et al., 2000), health problems (Sala et al., 2012) or repeated medical prescriptions (Petty et al., 2001). For Twitter linkage, in two of their reported studies, Mneimneh et al. (2020) measured the relevance of the data source through inquiring about the respondent's use of different social media platforms and the frequency of Twitter use among a Belgium student sample and a Saudi Arabian community sample.

Other individual level factors that have been discussed in the administrative data linkage literature include respondents' tendency to say 'yes' and their cognitive ability (Beninger et al., 2017; Sakshaug et al., 2012). While these could also be relevant to social media linkage, we focus this paper on the following set of factors: respondents' attitudes towards helping behaviour, general privacy concerns, privacy attitudes towards technology and relevance of social media to the respondent, as this set of individual-level factors has not been investigated in the same study among a probability online panel of Twitter users in the United States.

The second set of factors is related to design and include sample type (panel respondents or respondents of a cross-sectional study), the mode of the consent request (web administered, telephone or interviewer administered), consent language and framing, and consent placement within the survey instrument. Of these design factors, the one that has been found to be consistently related to consent to link administrative data with survey data is consent request placement. Several studies have found that placing the consent request earlier in the interview is associated with higher consent to link rates compared to placing the request towards the end of the interview (Kreuter et al., 2015; Sakshaug et al., 2013; Sakshaug & Vicari, 2018; Sala et al., 2014).

Several mechanisms have been proposed to explain this effect including: (1) signalling the importance of the linkage task by placing it at the beginning of the interview, (2) adopting a 'foot-in-the-door approach' (Freedman & Fraser, 1966) that speculates respondents to be more likely to comply with a subsequent request shortly after they have already agreed to an initial request, (3) leveraging the higher respondent engagement at the beginning of the interview (compared to the end of the interview) and (4) conveying the study sponsors to be more transparent and forthcoming in their request by placing it at the beginning rather than waiting until the end (after all the information has been collected). In spite of this consistent finding, many panel vendors raise concerns about asking their panellists to consent to link at the beginning of the interview and require researchers to place the consent request towards the end of the interview. This is likely to be driven by the perceived sensitive nature of the request and the related concern that placing the request at the beginning might lead respondents to break off or to refuse to be re-interviewed (if they are part of the panel).

This is the first study that empirically tests the effect of placing the consent request to link survey data with Twitter public data at the beginning of the interview compared to placing it at the end of the interview. Moreover, given that some respondents might have reservations for consenting at the beginning of the interview before they are aware of the full content of the survey questions, asking respondents who did not consent at the beginning to reconsider their decision at the end of the interview would allow respondents to make a more informed decision. This approach is also tested in this study for the first time, in an attempt to maximise the value of the linked data while adhering to the proper ethical conduct of social media research.

The third set of factors is social or environmental. While this study does not investigate any social and environmental factors, it is important to acknowledge the effect that these might play on respondents' decisions to consent to link their survey data with their social media data. Such factors have also been discussed in the administrative data linkage literature and include measures such as the number of consents that have been already given by other household members (Sala et al., 2012), a partner's or spouse's attitudes towards such consent requests, and the prevalence of fraud and data leaks (Beninger et al., 2017).

In light of the existing literature and the need to explore the properties of social media data for social science research, this study aims at investigating the following research questions to enrich the limited body of literature on this topic.

Research Question 1: What individual level characteristics, including attitudinal measures and social media engagement patterns are associated with providing consent to link survey data with Twitter data and reporting of a Twitter handle (controlling for respondent's socio-demographics)?
Research Question 2: Does the consent request placement—at the beginning versus the end of the interview—affect a respondent's decision to consent to link?
Research Question 3: Is there a significant gain in consent to link rates if respondents who did not consent at the beginning of the interview are re-asked to consider their consent decision at the end of the interview?
Research Question 4: How do these effects vary when considering not only consent to provide a handle but also the usability of the handle?

We answer these questions by analysing data that we collected through a web survey of Twitter users selected from an online probability panel that represents adults in the United States. Data were collected early in the first quarter of 2020 and included survey responses related to the main constructs identified above including attitudes towards helping behaviour, privacy concerns and social media engagement patterns. The sample was randomly divided into two groups, those who

received the consent to link request at the beginning of the survey, and others who received the request towards the end of the survey.

2 | METHODS

A random sample from Ipsos KnowledgePanel was used. KnowledgePanel is a probability-based online panel of US adults (18 years old and over) including those who reside in households without internet and who are provided a web-enabled device and free Internet service upon entry into the panel. Starting in 2009, panellists have been recruited through an Address-Based Sampling (ABS) technique. Details of the recruitment process are provided in 'KnowledgePanel: A Methodological Overview' report (link provided in the reference list, KnowledgePanel, 2021). On average, 11% of household members who are approached to join the panel complete the Core Profile survey, which covers demographics and household composition. The sample used for this paper was based on panellists who reported being Twitter users in 2019 and who provided information on their frequency of Twitter user. Based on this reported information, three strata were formed: daily users, weekly users, and those who use Twitter monthly or less than monthly. Daily users were oversampled and the differential probabilities of selection between the three strata were adjusted in the final sample by creating base weights defined as the inverse of the probability of selection in each stratum (further details are provided later when weights are discussed). The final completed number of interviews in each user stratum 923 (daily), 402 (weekly) and 379 (monthly or less), for a total of 1704 respondents.

The survey was fielded between 28 January 2020 and 20 February 2020 with a completion rate of 76%. An initial email was sent to all selected panel members inviting them to take a web survey through the Ipsos panel portal. Two reminders followed the initial invite. The first reminder was sent on the third day of the field period, and the second was sent out on the seventh day. All respondents were presented with a consent statement (consent request language found in Table A1) inquiring about their approval to share their Twitter handles with the research team for linking their survey data with their Twitter public information. Several Twitter informed consent language elements discussed in Bruer et al. (2021) were used. These include how the collected public Twitter information will be accessed, stored and used, the duration of Twitter data collection, and signalling the username with an '@' sign to clarify the piece of information needed. Those who consented to provide their Twitter handle received an incentive that is equivalent to $5.

All design and implementation features of the study were reviewed and approved by the Institutional Review Board of University of Michigan.

The questionnaire included 34 questions related to the respondent's use of social media, specific Twitter use behaviours, privacy concerns, attitudes towards helping behaviours, attitudes towards technology, racial prejudice, political knowledge and attitudes, political affiliation, religiosity, and a number of socio-demographics. The median interview length was about 6 min.

The design of the questionnaire included an experiment related to the position of the linkage consent request. A little under three quarter of the respondents (73.1%) were randomised to receive the consent request towards the beginning of the questionnaire and the remaining (26.9%) received the request towards the end of the questionnaire. Respondents who received the consent request at the beginning of the interview and who did not provide consent were re-asked towards the end of the interview. The number of respondents in each consent allocation group is presented in Figure 1.

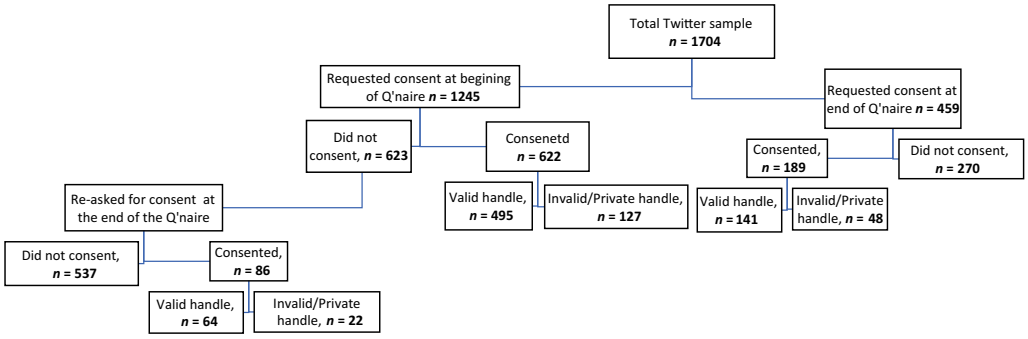

FIGURE 1 Distribution of Twitter sample by position of consent to link request and consent decision (consent request language in Table A1) [Colour figure can be viewed at wileyonlinelibrary.com]

Different sets of factors were identified in the introduction section and hypothesised to be correlated with a respondent's decision to consent to link their survey responses to their Twitter public data, which is the main outcome investigated and hereafter referred to as consent to link. Below is a description of these factors and their operationalisation. Further details on all the measures are provided in Table A1.

2.1 | Attitudes towards helping behaviour

This measure was constructed from two survey questions related to helping others in general and the importance of sharing possessions with others. An index summing up the respondent's answers to these two questions was created, and respondents were grouped into three categories based on the index score: low, medium and high.

2.2 | Privacy

Privacy concerns were measured through two survey questions, a general but direct privacy question, and an indirect indicator related to data access of online services. For general privacy concerns, respondents indicated whether they are very worried, somewhat worried, not very worried or not at all worried about their personal privacy. Only 1.4% of respondents indicated that they are 'not at all worried', so this category was grouped with the 'not very worried' in the analyses. For the online services question, respondents were asked whether they strongly agree, somewhat agree, somewhat disagree or strongly disagree that online services are more efficient because of increased access to personal data.

2.3 | Relevance of online data sources

This construct was measured through a few indicators. The first is an indirect indicator related to attitudes towards posting online videos operationalised through an agree/disagree question that was available as a frame variable (i.e. collected from panellists in a prior survey and available to researchers). The second indicator is specifically related to social media engagement. Respondents' engagement in social media was measured by adding the number of personal accounts the respondent identified on 10 social media platforms (Table A1).

The third group of indicators is related to the respondent's Twitter engagement pattern. To measure this, respondents were asked whether they use Twitter to read tweets, retweet, tweet about different topics or send messages. Respondents could choose multiple options. In addition, respondents indicated their frequency of Twitter use (daily, weekly, monthly or less than monthly).

2.4 | Consent request position

An indicator specifying whether the respondent received the request for consent to link towards the beginning or the end of the questionnaire was created.

2.5 | Reasons for not consenting to link

For a random sub-sample of 177 Twitter users, those who did not consent ($n = 82$) were asked an open-ended question about their non-consent reasons. All open responses were first reviewed and coded by one coder and then reviewed by the author.

2.6 | Twitter handles

All reported handles were checked manually to identify whether the handle existed on Twitter, belonged to a public figure, or was public. Accordingly, the handles were coded as usable or not usable. Among the unusable handles, 53% did not exist, 43% were related to private accounts, and 4% were related to non-personal accounts.

Finally in addition to the main factors described above, a number of respondent's socio-demographics were included as controls to provide a comparison to the existing literature on social media consent to link. Respondent's socio-demographics included age, gender, marital status, race/ethnicity, parental status, educational level, employment status, annual household income, household size, region, household type and political affiliation. The categories of these characteristics and their respective distributions are shown in Table 1. Many of these measures were already available as part of the panellists' profiles (i.e. frame variables) and were not collected in the current survey. The distinction between measures that originated from the frame versus survey measures (i.e. the current survey), the original categories and the recoded categories are provided in the Table A1.

All results presented in Tables 1–4 are weighted. Weights were created to account for the differential sampling rate within each Twitter use stratum and nonresponse and calibrated to represent the Twitter population of the full KnowlegePanel after correcting for misclassification of Twitter use frequency. Sampling weights were defined as the inverse of the probability of selection in each stratum. Non-response was adjusted through a response propensity model using available socio-demographic variables from the frame on all respondents and non-respondents. Variables included: age, gender, race/ethnicity, education, household size, head of the household status, housing type and ownership, household income, metro versus non-metro residence area, and region. Frequency of Twitter use was available on the sampling frame and used for stratification. An updated measure of Twitter use frequency was also collected during the survey. This allowed for estimating any measurement misclassification in the stratification variable (Jang et al., 2009).

TABLE 1 Socio-demographic characteristics of the Twitter sample ($n = 1646$)[a]

	Weighted percentage (unweighted, *n*)
Age (in years)	
18–29	35.4%(268)
30–44	32.3%(466)
45–59	19.8%(496)
60+	12.5%(416)
Gender	
Female	50.5%(721)
Male	49.5%(925)
Marital status	
Married/living with a partner	50.2%(1043)
Divorced/separated	8.1%(159)
Widowed	2.1%(46)
Never married	39.6%(398)
Race/ethnicity	
White, non-Hispanic	51.5%(1182)
Black, non-Hispanic	16.8%(153)
Hispanic	22.2%(191)
Others, non-Hispanic	9.5%(120)
Parent to a child currently in care of respondent	26.0%(422)
Education	
Less than high school	12.4%(41)
High school	28.5%(245)
Some college	26.2%(443)
At least a college degree	32.9%(917)
Employment status	
Employed	66.7%(1078)
Self-employed	6.2%(140)
Not working	15.6%(136)
Retired	7.2%(235)
Disabled	4.3%(57)
Annual household income	
<$40,000	31.7%(329)
$40,000–$74,999	25.2%(346)
$75,000–$149,999	21.4%(468)
≥$150,000	21.7%(503)

(Continues)

TABLE 1 (Continued)

	Weighted percentage (unweighted, *n*)
Household size	
One	17.2%(380)
Two-three	57.0%(881)
Four-five	21.2%(321)
Six or more	4.6%(64)
Region	
Northeast	14.3%(303)
Midwest	19.1%(403)
West	27.5%(352)
South	39.1%(588)
Housing type	
Detached house	63.4%(1140)
Attached house	7.2%(143)
Building with two or more apartments	26.6%(327)
Mobile home/Boat/RV	2.8%(36)
Political affiliation	
Strong Republican	11.7%(241)
Not strong Republican	18.9%(336)
Strong Democrat	22.2%(454)
Not strong Democrat	27.1%(421)
Independent	20.1%(194)

[a]The sample size dropped from 1704 to 1646 since the sample was restricted to respondents with no missing data on the variables used in this paper. There were 58 (3.4%) respondents who had missing data on at least one question. There was no single variable that had more than 1.1% missing data.

To correct for potential loss of sample efficiency because of this misclassification, an imputation model was used to predict an updated frame measure based on the observed frequency of use and demographics. The average of 10 imputations was calculated (Raghunathan et al., 2018). Any value great than or lower than six times the interquartile (from the median) was trimmed and used in the final weight.

Four weighted logistic regression models were used to investigate predictors of consent to link. The first two models predicted consent to link irrespective of whether the handle reported was usable or not (to answer research questions 1 and 2 about individual related predictors and consent request position effect). One of the models defined consenters as those respondents who consented to the initial request only (irrespective of the outcome of the second request for those who did not consent at the initial request). This model will be informative for studies that opt to ask for consent once. The second model define consenters as those who consented either at the initial request or upon the second request to investigate research question 3, related to gains in consent from a second request.

TABLE 2 Weighted consent rates and usability of the handle by the position of consent request

	Weighted % (unweighted, *n*)
Request towards the beginning of the questionnaire	
Consented at first request and useable handle	48.0%(487)
Consented at first request and unusable handle	10.6%(124)
Consented at second request and useable handle	4.4%(62)
Consented at second request and unusable Handle	1.0%(21)
Not consented at first and second request	36%(509)
Total	100.00%(1203)
Request towards the end of the questionnaire	
Consented and useable handle	30.7%(137)
Consented and unusable handle	18.3%(47)
Not consented	51.0%(259)
Total	100.0%(443)

The second set of models predicted consent to link and provide a useable handle to answer the fourth research question. In these two models, those who consented and provided an unusable handle were grouped with non-consenters. Again, the first of these two models only considered the outcome of the initial consent request, and the second model included the outcome of the initial request and the second request.

For all of the models, initially only socio-demographics were included as predictors. Of these only age, gender, marital status and education were found to be significantly associated with consent to link. These significant socio-demographics were then included in a second model with all other predictors related to helping behaviours, privacy concerns, relevance of online data sources and consent request position.

All analysis was conducted using SAS 9.4 software.

3 | RESULTS

3.1 | Description of the sample

Twitter respondents were split almost equally between males (49.5%) and females (50.5%). A little over a third were less than 30 years old, and about another third were between 30 and 44 years old. The oldest group (60 years and over) constituted 12.5% of the sample. About half were married and living with a partner, 39.6% were never married, and the remainder were previously married (either divorced, separated or widowed). A little over one quarter of the sample were parents who took care of a child under 18 years old at the time of the interview. A little over half of the weighted sampled identified themselves as white/non-Hispanic, 16.8% as black non-Hispanic and 22% as Hispanic.

In terms of socio-economics, about a third reported having at least a college degree, a little over a quarter reported some college education, 28.5% a high school degree, and the remaining

12.4% had not completed high school. The majority of respondents were either employed (66.7%) or self-employed (6.2) %, 15.6% were not working at the time of the interview and the remaining were either retired (7.2%) or disabled (4.3%). The sample was distributed across the different household income groups, with a little under a third making less than $40,000 annually, a quarter making between $40,000 and $74,999 and the rest equally split between the two highest income groups ($75,000–$149,999 and $150,000 or more). The majority of the sample had a medium-sized household, with 57% having two or three household members, and 21.2% four to five household members. Less than 5% had six or more members. The majority of the interviewed households lived in detached houses (63.4%), a little over a quarter lived in apartment buildings, and the rest lived in attached houses (7.2%) or mobile homes (2.8%).

Geographically, the sample was distributed among the four regions of the United States, with 39.1% living in the South, 19.1% in the Midwest, 27.5% in the West region, and 14.3% in the Northeast. About half of the sample identified themselves as Democrats (among which 22.2% strong Democrats), about 30% Republicans (among which 11.7% strong Republican), and 20.1% as Independents.

Complete distributions of these characteristics are provided in Table 1.

3.2 | Consent rates and useable handles

Table 2 provides a detailed distribution of respondents who consented to link and who provided a useable handle, grouped by the position of the consent request. Of the respondents who received the consent request towards the beginning of the questionnaire, 58.6% consented at the first request, of whom 48% provided a useable handle. An additional 5.4% consented at the second request, of whom 4.4% provided a useable handle. Of the respondents who received the request towards the end of the questionnaire, 49% consented, of whom 30.7% provided a useable handle.

Thus, even without re-asking respondents, placing the consent request towards the beginning of the interview compared to the end of the interview yields an additional 10% of respondents who consent to link (58.6% vs. 49%). This difference is larger, 17.3%, when considering only usable handles and increases to 21% when the second request is added.

3.3 | Individual-level characteristics and consent request position as predictors of consent to link

We first present predictors of consent to link at initial request (Table 3, Column A).

Among the socio-demographic controls, previously married respondents and Black respondents had higher odds of consenting to link upon initial request compared to never married respondents and White respondents (OR = 3.1 and OR = 2.76, respectively). Privacy concerns were one of the strongest predictors of consent to link. Respondents who reported not being worried about their privacy in general had higher odds of consenting to link than those who reported being very worried about privacy (OR = 2.57). In terms of social media engagement measures, the more personal accounts a respondent has on multiple platforms, the higher were the odds of consenting (OR = 1.21). Moreover, respondents who retweet were more likely to consent than those who did not (OR = 1.89). On the other hand, respondents who reported using Twitter daily were less likely to consent than those who reported using Twitter less than once a

TABLE 3 Odds ratio from logistic regression predicting consent to link and providing a Twitter handle among Twitter users ($n = 1646$)[a]

	A. Only initial request[b]	B. Initial and second request[c]
	OR (95% CI)	OR (95% CI)
Age (in years, ref. = 18–29)		
30–44	1.24 (0.74–2.09)	1.21 (0.72–2.02)
45–59	0.87 (0.51–1.48)	0.94 (0.55–1.63)
60+	0.79 (0.42–1.52)	0.73 (0.38–1.38)
Female (ref. = male)	0.83 (0.57–1.22)	0.89 (0.60–1.31)
Marital status (ref. = never married)		
Married/living with a partner	1.31 (0.84–2.06)	1.41 (0.89–2.23)
Previously Married	3.10 (1.58–6.10)**	3.49 (1.74–6.99)**
Race/ethnicity (ref. = White, non-Hispanic)		
Black, non-Hispanic	2.76 (1.46–5.21)**	2.46 (1.29–4.70)**
Hispanic	1.20 (0.71–2.03)	1.09 (0.65–1.84)
Others, non-Hispanic	0.89 (0.47–1.68)	0.86 (0.45–1.64)
Education (ref. = less than high school)		
High school	1.09 (0.44–2.72)	1.42 (0.57–3.53)
Some college	1.60 (0.64–3.99)	1.77 (0.72–4.39)
At least college degree	0.72 (0.30–1.74)	0.90 (0.37–2.19)
Attitudes towards helping behaviours (ref. = low)		
Medium	0.90 (0.49–1.65)	1.04 (0.56–1.93)
High	1.22 (0.58–2.60)	1.54 (0.70–3.37)
Worried about privacy (ref. = very)		
Somewhat worried	1.24 (0.77–1.99)	1.13 (0.70–1.80)
Not worried	2.57 (1.47–4.50)**	2.23 (1.27–3.91)**
Online Services efficient b/c access to personal data (ref. = strongly disagree)		
Somewhat disagree	0.79 (0.46–1.36)	0.91 (0.54–1.55)
Somewhat agree	1.00 (0.58–1.74)	1.05 (0.61–1.83)
Strongly agree	1.89 (0.93–3.86)	1.93 (0.94–3.93)
Online video posting (ref. = do not agree)		
Somewhat agree	1.44 (0.86–2.41)	1.60 (0.95–2.71)
Agree	0.62 (0.30–1.28)	0.95 (0.43–2.09)
Strongly agree	1.79 (0.71–4.50)	1.49 (0.57–3.88)

(Continues)

TABLE 3 (Continued)

	A. Only initial request[b]	B. Initial and second request[c]
	OR (95% CI)	OR (95% CI)
Number of social media platforms used	1.21 (1.08–1.35)**	1.19 (1.06–1.33)**
Type of Twitter use		
None	0.34 (0.11–1.01)	0.40 (0.13–1.21)
Read Tweet	0.60 (0.29–1.21)	0.59 (0.29–1.22)
Retweet	1.89 (1.19–3.00)**	2.10 (1.32–3.35)**
Tweet	0.88 (0.56–1.40)	0.89 (0.56–1.41)
Send messages	1.07 (0.62–1.85)	0.86 (0.51–1.46)
Frequency of Twitter use (ref ≤monthly)		
Daily	0.57 (0.33–0.97)*	0.56 (0.33–0.96)*
Weekly	0.74 (0.42–1.28)	0.78 (0.44–1.39)
Monthly	1.14 (0.61–2.12)	1.02 (0.54–1.92)
Consent request beginning versus End of Q	1.45 (0.93–2.27)	1.87 (1.20–2.91)**

[a] The sample size dropped from 1704 to 1646 since the sample was restricted to respondents with no missing data on the variables used in this paper.

[b] Respondents who were randomised to receive the consent request towards the beginning of the interview were asked for consent twice. The first time was at the beginning of the interview, and if they did not consent, then they were asked to reconsider their consent decision towards the end of the interview (i.e. second request). This column only considers those who consented at the first request.

[c] This column includes both those who consented upon the first or the second request.

*p-value < 0.05.

**p-value < 0.01.

month (OR = 0.57). Finally, the index measuring attitudes towards helping behaviours was not significantly associated with consent to link.

3.4 | Value of adding a second later request when the initial consent is positioned at the beginning of the interview

When combining respondents who consented upon the second request with those who consented to the first request and comparing them to non-consenters (Table 3, column B), the associations between marital status, race/ethnicity, social media engagement (higher number of social media accounts on different platforms and retweeting) and consenting to link remain. For privacy concerns, respondents who reported not being worried about their privacy compared to those who reported being very worried are still more likely to consent (OR = 2.23, Table 3, column B), albeit with a lower odds ratio compared to the model that only considers initial consenters (OR = 2.57, Table 3, column A). Moreover, placing the consent request towards the beginning of the questionnaire, followed by asking non-consenters to reconsider towards the end of the interview increased the odds of consenting to link by 1.87 times compared to placing the consent towards the end of the interview (with no second request), making the effect of consent placement significant.

3.5 | Individual-level characteristics and consent position as predictors of consent to link and provide a useable handle

Similar to general consenters (those who consented to link and provided a handle irrespective of its usability), previously married respondents (compared to never married) and Black respondents (compared to White) were more likely to consent and provide a usable handle upon the first request (Table 3, column A; OR = 2.75 and OR = 2.22). However, female respondents had lower odds of consenting with a usable handle compared to males (OR = 0.63, Table 4, column A).

Once again, the number of personal accounts a respondent has on different platforms, and retweeting (compared to not retweeting) were also positively associated with consenting to link and providing a useable handle (OR = 1.16 and OR = 1.73, respectively). Attitudes towards privacy were also among the strongest predictors. Respondents who reported not being worried about their privacy had higher odds (OR = 2.86) of consenting and providing a usable handle than those who reported being very worried. Placing the consent request at the beginning of the interview, even when only considering those who consented and provided a usable handle upon the first request, increased the odds of consenting by about two times (OR = 2.07).

When considering both those respondents who consented at the first request or those who consented at the second request (Table 4, column B), the effect of the request position increases to an odds ratio of 2.48. Moreover, age emerges as another predictor. Specifically, older respondents (at least 60 years) were less likely to consent and provide a useable handle compared to respondents who were 18–29 years old (OR = 0.43).

3.6 | What are the reasons provided by respondents for not consenting?

A random sub-sample of respondents who did not consent to link were asked to state their reasons in an open-ended format ($n = 82$). The most commonly reported reason was related to privacy and personal information (56%). Examples of reasons given include: '*Didn't want my info out there and possibly sold or compromised*'; '*I don't want give any information out*'; and '*I'm not comfortable sharing all information, particularly personal information*'. About 11% of respondents reported a general preference to not sharing this information such as: '*I just don't think you need to know it*'; '*Just prefer not to*.' Another 10% reported lack of Twitter use as a reason such as: '*I don't use it enough for it to be helpful so you don't need my information*'. Finally, about 7% reported reasons related to the language of the request. For example, '*Even though it is public information, something about it just doesn't feel right*'; '*If the [incentive] information was displayed more prominently, I would probably have said yes*'.

4 | DISCUSSION

This study contributes to the emerging literature on consent to link Twitter public data to survey data through its four unique design features. First, the study investigates a range of individual-level characteristics beyond demographics among a probability panel of Twitter users in the United States. The factors include attitudes towards helpful behaviours, privacy concerns and the relevance of social media to respondents. Second, this is the first study that randomly assigns respondents to receive a Twitter consent to link request either at the beginning of the interview or towards the end of the interview and evaluates the position effect on consent rates.

TABLE 4 Odds ratio from logistic regression predicting consent to link and provide a usable handle among Twitter users ($n = 1646$)[a]

	A. Only initial request[b]	B. Initial and second request[c]
	OR (95% CI)	OR (95% CI)
Age (in years, ref. = 18–29)		
30–44	1.16 (0.67–2.02)	1.10 (0.64–1.89)
45–59	0.83 (0.47–1.46)	0.88 (0.50–1.56)
60+	0.50 (0.25–1.01)	0.43 (0.22–0.85)*
Female (ref. = male)	0.63 (0.42–0.94)*	0.64 (0.43–0.95)*
Marital status (ref. = never married)		
Married/living with a partner	1.43 (0.88–2.32)	1.51 (0.94–2.43)
Previously married	2.75 (1.32–5.70)**	2.81 (1.30–6.08)**
Race/Ethnicity (ref. = White, Non-Hispanic)		
Black, non-Hispanic	2.22 (1.16–4.23)*	1.88 (1.01–3.52)*
Hispanic	1.48 (0.86–2.55)	1.30 (0.76–2.23)
Others, non-Hispanic	0.71 (0.36–1.41)	0.66 (0.34–1.28)
Education (ref. = less than high school)		
High school	0.79 (0.32–1.96)	1.04 (0.41–2.63)
Some college	1.08 (0.44–2.68)	1.23 (0.50–3.04)
At least college degree	0.55 (0.23–1.28)	0.69 (0.29–1.63)
Attitudes towards helping behaviours (ref. = low)		
Medium	0.95 (0.48–1.88)	1.15 (0.58–2.26)
High	1.15 (0.49–2.72)	1.44 (0.62–3.34)
Worried about privacy (ref. = very)		
Somewhat worried	1.25 (0.75–2.08)	1.10 (0.68–1.77)
Not worried	2.86 (1.57–5.21)**	2.29 (1.28–4.10)**
Online Services efficient b/c access to personal data (ref. = strongly disagree)		
Somewhat disagree	1.34 (0.72–2.47)	1.51 (0.83–2.77)
Somewhat agree	1.32 (0.72–2.42)	1.32 (0.72–2.42)
Strongly agree	1.37 (0.63–2.99)	1.44 (0.67–3.10)
Online video posting (ref. = Do not Agree)		
Somewhat agree	1.25 (0.73–2.13)	1.34 (0.78–2.30)
Agree	0.61 (0.31–1.19)	0.80 (0.40–1.60)
Strongly agree	1.67 (0.62–4.51)	1.40 (0.52–3.77)
Number of social media platforms used	1.16 (1.04–1.29)**	1.14 (1.02–1.28)*

(Continues)

TABLE 4 (Continued)

	A. Only initial request[b]	B. Initial and second request[c]
Type of Twitter use		
None	0.48 (0.14–1.66)	0.59 (0.18–2.00)
Read Tweet	0.99 (0.48–2.03)	0.97 (0.47–1.99)
Retweet	1.73 (1.09–2.76)*	1.91 (1.21–3.02)**
Tweet	1.24 (0.78–1.97)	1.26 (0.79–2.01)
Send messages	0.97 (0.57–1.65)	0.78 (0.46–1.32)
Frequency of Twitter use (ref ≤monthly)		
Daily	0.59 (0.34–1.04)	0.58 (0.34–1.01)
Weekly	0.72 (0.41–1.27)	0.70 (0.40–1.21)
Monthly	0.88 (0.47–1.65)	0.79 (0.42–1.47)
Consent request beginning versus end of Q	2.07 (1.26–3.39)**	2.48 (1.54–3.99)**

[a] The sample size dropped from 1704 to 1646 since the sample was restricted to respondents with no missing data on the variables used in this paper.

[b] Respondents who were randomised to receive the consent request towards the beginning of the interview were asked for consent twice. The first time was at the beginning of the interview, and if they did not consent then they were asked to reconsider their consent decision towards the end of the interview (i.e. second request). This column only considers those who consented at the first request.

[c] This column includes both those who consented upon the first or the second request.

*p-value < 0.05;

**p-value < 0.01.

Third, the study investigates the added benefit of asking respondents who initially declined consent at the beginning of the interview to reconsider their consent decision towards the end of the interview after they had the chance to review the survey content. Fourth, given that a proportion of consenters will report a non-useable handle, the study investigates whether the factors associated with consenting and reporting a useable handle are the same as those associated with general consent to link (irrespective of the usability of the handle).

The consent to link rates found in our study are in between the rates reported in the literature. While Wojcik and Hughes (2019) reported a consent to link rate of 90% among a US online probability sample, all other reported rates using a web administered mode were between 23% and 27% (among panel members or respondents of a cross-sectional sample). Our consent to link rates upon initial request is in the range of 49%–58% depending on whether the request was placed towards the beginning of the interview or towards the end. The nine percentage point gain in consent to link rate when the request is placed at the beginning of the interview is similar to the gain reported in the administrative linkage literature (Sakshaug et al., 2013; Sakshaug & Vicari, 2018; Sala et al., 2014). The gain in consent rate that is attributed to placing the request at the beginning of the interview would be higher if respondents who did not consent at the beginning are asked to reconsider their decision at the end of the interview after they have the chance to review the content of the survey. When considering both initial and second requests, placing the consent request towards the beginning of the interview and re-asking towards the end increases consent to link rate by 15% compared to asking respondents only once at the end of the interview. This upfront approach followed by a re-ask at the end of the interview maximises the benefit of the different decision consent mechanisms that might be relevant to some respondents more than others. By placing the request at the beginning rather than at the end, some respondent would

be more agreeable to consent right after they have decided to cooperate with the survey itself, others might feel more engaged at the beginning of the survey, and some might appreciate the transparency and forthcoming approach of the sponsor. Yet, placing the request at the beginning might be considered too sensitive for some respondents, especially when they are still not sure about the content of the questionnaire. These respondents might not provide their consent if the request is placed at the beginning. By re-asking about their permission to link, these respondents are given a second chance and can make a more informed decision based on the content of the survey while respecting their continued decision to decline the request.

After consenting to link, and unless respondents are asked to log in directly to their social media account, most studies request respondents to self-report their username or handle. In our study, the linkable consent rate drops by 11.6%–18.3% (depending on the position of the request) when unusable handles are removed (either because they did not exist, they are private, or they belonged to a public figure or celebrity). These rates are similar to Hughes et al. (2020), who found a 14 percentage point decline in consent rate when non-valid usernames are removed. After removing such unusable handles, the difference in consent rate based on the position of the request is 21% in favour of placing the request at the beginning followed by a re-ask for non-consenters at the end of the interview. Bruer et al. (2021) provide a good discussion about informed consent for different platforms and the issue of private versus public accounts which impacts usable consent rates.

In adopting an initial upfront request to link with a backend re-ask, about half of the respondents consented to link and provided a usable handle. Those who did are generally not worried about their privacy, have a higher number of social media accounts, and have indicated that they retweet (compared to those who did not consent or who did not provide a useable handle). The association between privacy concerns and consent to link is not surprising and has been found in the administrative linkage literature albeit through indirect measures of privacy concerns such as refusal to answer sensitive items (Jenkins et al., 2006; Sala et al., 2012), resistance to participate in earlier waves of a panel study, or through interviewer ratings of a respondent's uncooperative behaviour (Sakshaug et al., 2012). In the Twitter literature specifically, when Fiesler and Proferes (2018) inquired about Twitter users' attitudes towards using their public Twitter data for academic research (even without linkage to other sources of data), privacy concerns were commonly reported by users. In our study, privacy and the personal nature of Twitter data was also the most common reason given by respondents when asked about their reasons for not consenting. While our approach of giving respondents the chance to be informed about the content of the questionnaire before consenting and re-asking them to consider their decision towards the end of the questionnaire did not completely alleviate the privacy concerns of many respondents, it did seem to diversity the pool of consenters. Consenters who re-considered their decision upon the second request had greater privacy concerns than those who consented at the initial request. This is reflected by a reduction in the effect of privacy concerns from model A (OR = 2.86) which only considered initial consenters to model B (OR = 2.29) which considered initial and secondary consenters.

In addition to privacy concerns, the relevance of social media to respondents was also associated with consent to link. The higher the number of social media platforms used by respondents, the higher the odds to consent to link. A similar finding has been reported in a study among undergraduate students in Belgium who were also requested to consent to link their survey data with their Twitter public information (Mneimneh et al., 2020). Moreover, respondents who reported being active users of Twitter through retweeting compared to inactive Twitter users were also more likely to consent to link. Though respondents who reported Tweeting were slightly more

likely to consent compared to inactive users, the effect was not statistically significant. The data source relevance effect has also been observed in the administrative literature, where respondents who have indicated having health problems or who reported having multiple medical prescriptions were found more likely to consent to link their data to their health records (Petty et al., 2001; Sala et al., 2012; Woolf et al., 2000).

The third individual-level factor we investigated related to respondent's attitudes towards helping others. Our index, comprised of a general helping statement and a statement about the importance of sharing possessions with others, was not found to be associated with consent to link. While this could be due to the nature of the measure itself, in another study that also investigated consent to link survey data to Twitter data among an Address-Based Sample (ABS) of the US population, one helping behaviour factor was not found to be related to consent and the other was found to be marginally related (Mneimneh et al., 2020).

Finally, two important effects that emerged after unusable handles were removed are related to the age and gender of respondents. Older respondents (60 years and above) and females were less likely to consent and provide a usable handle compared to younger respondents (18–29 years) and males, respectively. Upon investigating the type of unusable handle, older respondents reported higher rates of handles that did not exist compared to younger respondents, and females reported higher rates of private handles than males. Thus, the former age effect might be attributed to difficulty in recalling a usable handle while the latter gender effect might be attributed to females generally preferring to have private Twitter accounts compared to males. When only considering consent to link (irrespective of the usability of the handle) the magnitude of age and gender effects were lower and not significant.

A few considerations are worth nothing with respect to our study design and findings. One of the main concerns of placing the consent to link request at the beginning is its potential effect on respondents' answers to the subsequent questions, creating context effects. Given that we randomised the position of the request, we are able to test for such effects. The biggest concern is related to the privacy-related questions and attitudes towards online posting. The response distributions of these measures did not differ between the two consent placement positions. The other concern is the potential confounding between the consent decision and the answers to the helpful attitudes questions. For example, a respondent might report they consider it important to be helpful because they just consented. However, in our study, we found neither a direct effect of helping attitudes on the consent decision nor an interaction effect with the position of the request. While our helping attitude measures were not associated with the decision to consent to link, given the limited research on social media linkage, researchers are encouraged to continue investigating the association between other measures of engagement in helpful behaviour, helpful attitudes and consent to link. Third, this study only investigates one design feature (consent position); other design features including consent phrasing which can potentially lower respondents' privacy concerns, higher incentives and other methods of administration might also play an important role in a respondent's decision to consent and would be valuable for future investigations. Finally, while the demographic composition of Twitter users in this study is generally similar to a Pew Research study using a nationally representative sample of US adult Twitter users (Wojcik & Hughes, 2019), Twitter users have some demographics differences from the general adult population and non-Twitter users. Twitter users were found to be younger, more educated, have higher income and more likely to be democrats than the general public (Wojcik & Hughes, 2019) and non-Twitter users (Mneimneh, 2021). Such differences are compounded when considering the sub-sample of Twitter users who consented to link. Thus, research findings based on Twitter user samples or consented Twitter users might not generalise to the US adult population,

especially for measures associated with age, education, income, political affiliation or privacy attitudes.

To conclude, there is still a lot of uncertainty about the properties of social media data and their utility for social science research. Linking social media data to other data sources (such as administrative data or survey data) have the potential to shed light on the properties of social media data itself. When linkage is done at an individual level, informed consent is needed, offering the opportunity for respondents to decline. The value of the linked data, therefore, resides in maximising consent to link rates while adhering to the proper ethical research practices.

The findings of our study reveal that respondents' level of privacy concerns and their social media engagement patterns are associated with the decision to consent to link and to provide a usable handle. Respondents who are less concerned about their privacy, those who have accounts on multiple social media platforms and who retweet are more likely to consent to link. Most importantly, designing the placement of the consent request at the beginning of the interview, with a follow up among non-consenters towards the end of the interview, compared to a single request at the end of the interview, increases the odds of consenting and provide a useable handle by about two and a half times, and diversifies the pool of consenters by recruiting more respondents who indicate a higher level of concern. Such findings can guide the design of future studies interested in requesting consent to link survey data with Twitter data. However, researchers need to consider that these findings are derived from a study among a sample of Twitter users in the United States who are part of an online probability panel. Consent rates among other types of samples such as a cross-sectional samples are expected to be lower.

FUNDING INFORMATION

This research was funded by the National Science Foundation, Division of Social and Economic Sciences, grant number 1259985, and an internal grant from the Michigan Institute for Data Science and Data Acquisition for Data Science, University of Michigan.

DATA AVAILABILITY STATEMENT

Research data have not been shared at this time yet.

ORCID

Zeina Mneimneh https://orcid.org/0000-0002-1091-7838

REFERENCES

Al Baghal, T., Sloan, L., Jessop, C., Williams, M.L. & Burnap, P. (2019) Linking Twitter and survey data: the impact of survey mode and demographics on consent rates across three UK studies. *Social Science Computer Review*, 38(5), 517–532.

Benton, A., Coppersmith, G. & Dredze, M. (2017) Ethical research protocols for social media health research. *Proceedings of the 1st ACL workshop on ethics in natural language processing*, pp. 94-102.

Bruer, J., Baghal, T.A., Sloan, L., Bishop, L., Kondyli, D. & Linardis, A. (2021) Informed consent for linking survey and social media data: differences between platforms and data types. *IASSIST Quarterly*, 45(1), 1–27. https://doi.org/10.29173/iq988

Beninger, K., Digby, A., Dillon, G. & McGregor, J. (2017). Understanding Society: How people decide whether to give consent to link their administrative and survey data. Working paper Series No. 2017-13. ESRC Economic & Social Research Council.

Fiesler, C. & Proferes, N. (2018) "Participant" perceptions of Twitter research ethics. *Social Media + Society*, 4(1), 2056305118763366.

Freedman, J.L. & Fraser, S.C. (1966). Compliance without pressure: The foot-in-the-door technique. *Journal of Personality and Social Psychology*, 4(2), 195–202. https://doi.org/10.1037/h0023552

Henderson, M., Jiang, K., Johnson, M. & Porter, L. (2019) Measuring Twitter use: validating survey-based measures. *Social Science Computer Review*, 39(6), 0894439319896244.

Hughes, A., Remy, E., Shah, A., McCabe, S., Lazer, D. & Hobbs, W. (2020) A voval minority: assessing the representativeness of tweeters and tweets. *BigSurv 2020: big data meets survey science*, *virtual conference*.

Internet Live Stats. (2021) https://www.internetlivestats.com.

Jang, D., Sukasih, A., Lin, X., Kang, K.H. & Cohen, S.H. (2009) Effects of misclassification of race/ethnicity categories in sampling stratification on survey estimates. In: *Proceedings of the American statistical association, survey methods section*. Alexandria: American Statistical Association, pp. 3414–3428.

Jenkins, S.P., Cappellari, L., Lynn, P., Jäckle, A. & Sala, E. (2006) Patterns of consent: evidence from a general household survey. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(4), 701–722.

KnowledgePanel. KnowledgePanel: "A Methodological Overview". Available at: https://www.ipsos.com/sites/default/files/ipsosknowledgepanelmethodology.pdf [Accessed 3rd March 2021].

Kreuter, F., Sakshaug, J.W., Schmucker, A., Singer, E. & Couper, M.P. (2015) *Privacy, data linkage, and informed consent*. Proceedings of the 70th Annual Conference of the American Association for Public Opinion Research, Hollywood, FL.

Lane, J., Stodden, V., Bender, S. & Nissenbaum, H. (Eds.). (2014) *Privacy, big data, and the public good: frameworks for engagement*. New York, NY: Cambridge University Press.

Mneimneh, Z.N. (2021) *Presentation at the American association for public opinion research annual conference*, Virtual.

Mneimneh, Z.N., McClain, C., Bruffaerts, R. & Altwaijri, Y.A. (2020) Evaluating survey consent to social media linkage in three international health surveys. *Research in Social and Administrative Pharmacy*, 17(6), 1091–1100. https://doi.org/10.1016/j.sapharm.2020.08.007.

Petty, D.R., Zermansky, A.G., Raynor, D.K., Lowe, C.J., Vail, A. & Freemantle, N. (2001) Clinical medication review by a pharmacist of patients on repeat prescriptions in general practice. *International Journal of Pharmacy Practice*, 9(S1), 47.

Raghunathan, T., Berglund, P.A. & Solenberger, P.W. (2018) *Multiple imputation in practice: with examples using IVEware*. Boca Raton: CRC Press.

Sakshaug, J.W. & Vicari, B.J. (2018) Obtaining record linkage consent from establishments: the impact of question placement on consent rates and bias. *Journal of Survey Statistics and Methodology*, 6(1), 46–71.

Sakshaug, J.W., Couper, M.P., Ofstedal, M.B. & Weir, D.R. (2012) Linking survey and administrative records: mechanisms of consent. *Sociological Methods & Research*, 41(4), 535–569.

Sakshaug, J., Tutz, V. & Kreuter, F. (2013) Placement, wording, and interviewers: identifying correlates of consent to link survey and administrative data. *Survey Research Methods*, 7(2), 133–144.

Sala, E., Burton, J. & Knies, G. (2012) Correlates of obtaining informed consent to data linkage: respondent, interview, and interviewer characteristics. *Sociological Methods & Research*, 41(3), 414–439.

Sala, E., Knies, G. & Burton, J. (2014) Propensity to consent to data linkage: experimental evidence on the role of three survey design features in a UK longitudinal panel. *International Journal of Social Research Methodology*, 17(5), 455–473.

Sloan, L., Jessop, C., Al Baghal, T. & Williams, M. (2020) Linking survey and twitter data: informed consent, disclosure, security, and archiving. *Journal of Empirical Research on Human Research Ethics*, 15(1-2), 63–76.

Wojcik, S. & Hughes, A. (2019) *Sizing up Twitter users*. Pew Research Center. Available from: https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/ [Accessed 6th April 2020]

Woolf, S.H., Rothemich, S.F., Johnson, R.E. & Marsland, D.W. (2000) Selection bias from requiring patients to give consent to examine data for health services research. *Archives of Family Medicine*, 9(10), 1111.

Zimmer, M. (2010) "But the data is already public": on the ethics of research in Facebook. *Ethics and Information Technology*, 12(4), 313.

How to cite this article: Mneimneh, Z. (2022) Evaluation of consent to link Twitter data to survey data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 185(Suppl. 2), S364–S386. Available from: https://doi.org/10.1111/rssa.12949

APPENDIX

TABLE A1 Measures and variables used in analysis

Measures	Available from frame/survey question	Variable used in analysis
Helping behaviour	Two Survey Questions: On a scale of 1 to 5, where 1 means 'not at all' and 5 means 'completely', how well does each of the following statements describe you? Please select a number between 1 and 5. Q1. I consider it important to share my possessions with others: 1(Does not describe me at all), 2, 3, 4, 5 (Describes me completely) Q2. I consider it important to try to help others. Same scale above.	Summed responses on questions Q1 and Q2 and recoded into: Low helping behaviour = sum between 1 and 4 Medium helping behaviour = sum between 5 and 8 High helping behaviour = sum greater than 9
Privacy		
General privacy concern	Survey Question: In general, how worried are you about your personal privacy? Very worried, Somewhat worried, Not very worried, Not worried at all	Recoded original categories into: Very worried, Somewhat worried, Not worried (collapsing not very worried and not worried at all)
Privacy concerns of online services	Survey Question: I appreciate that online services are more efficient because of the increased access they have to my personal data: Strongly agree, Somewhat agree, Somewhat disagree, Strongly disagree	Kept original coding
Relevance of online data source to the respondent		
Attitudes towards online video posting	Available on the frame using the following question: Do you agree or disagree with the following statements?: I like to post online video content that I create (such as to YouTube): Do not agree, Somewhat agree, Agree, Strongly agree	Kept original coding
Number of social media platforms used	Survey Question: Do you currently have a personal account on any of the following? (select all that apply): Facebook, Twitter, Instagram, Snapchat, Reddit, Pinterest, Tumblr, Youtube, LinkedIn, WhatsApp, None of the above	Summed up into an index from 1 to 10

(Continues)

TABLE A1 (Continued)

Measures	Available from frame/survey question	Variable used in analysis
Type of Twitter use	Survey Question: Think of your current use of your personal Twitter account. Generally speaking, do you use this account to: Read tweets from people, organisations, and/or users you follow; retweet messages from people, organisations, and/or users you follow; tweet about yourself, your family, and/or your friends; tweet about news or current events; tweet about work-related topics; send direct messages; other, please specify	Recoded into: Do not use (based on open ends); Read Tweets Retweet; Tweet of any topic; Send direct messages
Frequency of Twitter use	Survey Question: Considering all the ways you use Twitter on your personal account, in an average month, how often do you use Twitter with a computer, tablet, cell phone, or an app? At least once a day; At least once a week but less than daily; At; least once a month but less than weekly; Less than once a month	Kept original coding
Consent request language	Initial ask either at the beginning or at the end of the questionnaire using the following phrasing 'As part of this project, we would like to understand how survey responses relate to social media use. To help us explore this, we would like to ask for your permission to collect your public Twitter information and analyse it for current and future research purposes only. Information will be collected over a 12-month period using automated computer programs. Your consent is completely voluntary, and your information will be kept confidential and stored in a secure database. As a token of our appreciation, those who give us permission will receive 5000 points. Do we have your permission to collect the public information from your personal Twitter account(s)?' Re-ask for consent for respondents who received the initial request at the beginning using the following phrasing. 'Earlier in the survey, you indicated that you did not want to provide us with your Twitter handle. We would like to remind you that the public information from your Twitter account would only be used for current and future research purposes, and that as a token of our appreciation, those who give us permission will receive 5000 points. Would you like to reconsider your decision?'	Created a binary indicator (consent request at beginning vs. at end)

(Continues)

TABLE A1 (Continued)

Measures	Available from frame/survey question	Variable used in analysis
Reason for not consenting	Open-ended survey question: Thank you. We are interested in how we can better design our requests in the future. Can you tell us why you did not consent to give us your Twitter information?	Open responses were first reviewed and coded by one coder and then reviewed by the author
Socio-demographics		
Age in years	Available on the frame as open ended	Recoded into: 18–29; 30–44;45–59; ≥60
Gender	Available on the frame: Female, Male	Kept original coding
Marital status	Survey Question: Are you currently married, separated, divorced, widowed, never married, or living with a partner	Recoded into: married/living with partner; previously married, single (never married)
Race/Ethnicity	Available on the frame: White/non-Hispanic, Black/non-Hispanic, Others/non-Hispanic, Hispanic, two races/non-Hispanic	Recoded into: White/non-Hispanic, Black/non-Hispanic; Hispanic; Others/non-Hispanic (including two races/non-Hispanic)
Parent to a child currently taking care of	Survey Question: Are you a parent to any biological, adopted, foster, or step child that currently lives with you and who is 17 years old or under? Yes/No	Kept original coding
Educational level	Available on the frame as: less than high school, high school, some college, at least a college degree	Kept original coding
Employment status	Available on the frame: employed, self-employed, not working (laid off), looking for work, retired, disabled, others	Recoded into: employed; self-employed; not working/others; retired; disabled
Annual household income	21 categories starting with <5000 USD to more than 250,000 USD	Recoded into: <40,000; 40,000–74,999; 75,000-149,999; ≥150
Household size	Available on the frame as open ended with a range of 1–13	Recoded into: One, two-three, four-five, six or more
Region	Available on frame as: Northeast, Midwest, West, and South	Kept original coding
Housing type	Available on frame as: detached house, attached house, building with two or more apartments, mobile home, boat, RV	Recoded into: detached house; attached house; building with two or more apartments; others
Political affiliation	Three survey questions: Q1. Generally speaking, do you usually think of yourself as a Republican, a Democrat, an Independent, or what? Q2. Would you call yourself a strong (Republican/Democrat) or a not very strong (Republican/Democrat)? Q3. Do you think of yourself as closer to the Republican Party or to the Democratic Party?	Answers from the three questions were recoded into: strong Republican; not strong Republican; strong Democrat; not strong Democrat; Independent