Test-retest reliability of four U.S. non-probability sample sources

Mario Callegaro, Inna Tsirlin, Yongwei Yang & Qiao Ma
Google

77th AAPOR Conference
Chicago: May 11-13 2022

AAPOR
AMERICAN ASSOCIATION FOR PUBLIC OPINION RESEARCH
77th Annual Conference
May 11-13, 2022
Sheraton Grand Chicago • Chicago, IL
www.aapor.org
#aapor
Come Together
Advancing Inclusion and Equity
Through Data Collection,
Measurement, and Community

Motivation for this research

In User Experience and Market Research is very common to set up survey trackers (and in polling too!)

- Trackers are cross sectional surveys measured on independent samples from the same population at different points in time

- **What happens when you use non probability online panels for a tracker?**

Our goal

Compare the stability of estimates from non probability online panels over a short time period

- **Assumption:** General attitudes should not change in two weeks
  (unless major news impact)

- **We are not looking at accuracy of the estimates** when compared with a benchmark

Previous inspiring study (2020)

CAN WE COUNT ON YOU?

ASSESSING THE RELIABILITY AND VALIDITY OF PANELS AROUND THE WORLD

maru/BLUE

14 countries:
France, Italy, Spain, Germany, Russia, South Africa, Thailand, China, Australia, Singapore, India, Australia, Brazil
2 non probability online panels per country (28 in total)
Study repeated twice one week or so apart
500 respondents per panel per wave

Benchmark study + test retest study

“In total, there were 16 [*out of 28*] suppliers in 9 countries that got a score of between 85% and 100% on both reliability and validity” (Page 4)

What did we do?

We simultaneously ran identical surveys (N=1,500 each) on 4 popular survey platforms in the U.S.

To measure consistency, we repeated the survey two weeks after the original run
(Last week of June - second week of July 2021)

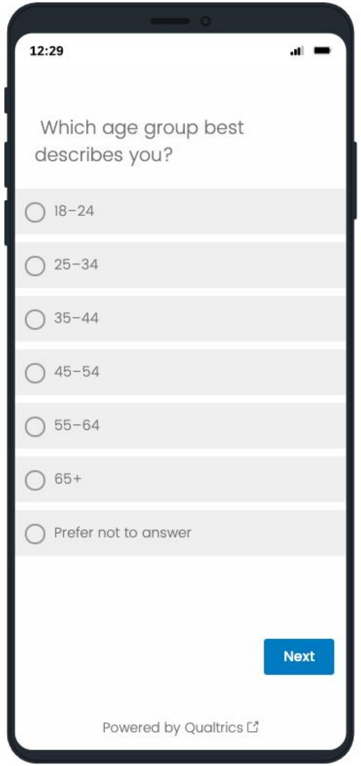
Data were weighting by age and gender using the 2019 Gallup World Poll, US, 18+ general online population.

Online panels:

- Google surveys on Publisher Network (GCS on PN) - *unpaid* panel of people browsing Publisher Network websites, their content is blocked by the surveys. Survey built using GS surveys engine (only 10 questions allowed).
- Google surveys on GOR (GCS on GOR) - paid panel of people who installed the Google Opinion Rewards app. Survey built using GS surveys engine (only 10 questions allowed).
- Qualtrics - paid panels surveyed by Qualtrics using Mfour (90%) as main source + Cint (10%)*.
- Amazon mTurk - paid *panel* of people signed up to do jobs on mTurk, survey built using the Qualtrics survey engine. mTurkeres were selected as US-only workers, at least 1 previously approved study (HIT), and 90% approval rate.

Questions used in the surveys

Demo: Age



Which age group best describes you?

○ 18–24

○ 25–34

○ 35–44

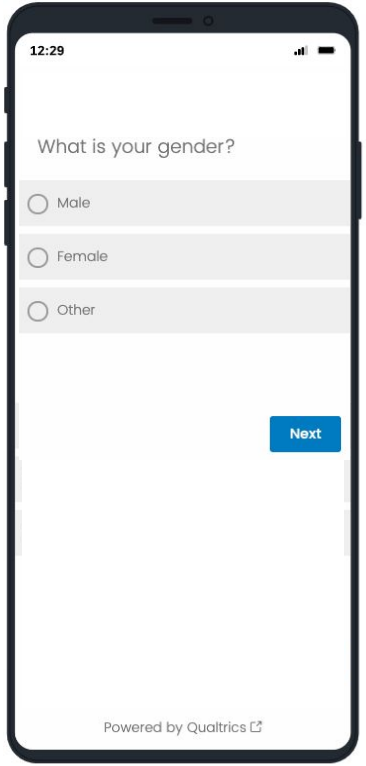
○ 45–54

○ 55–64

○ 65+

○ Prefer not to answer

Demo: Gender



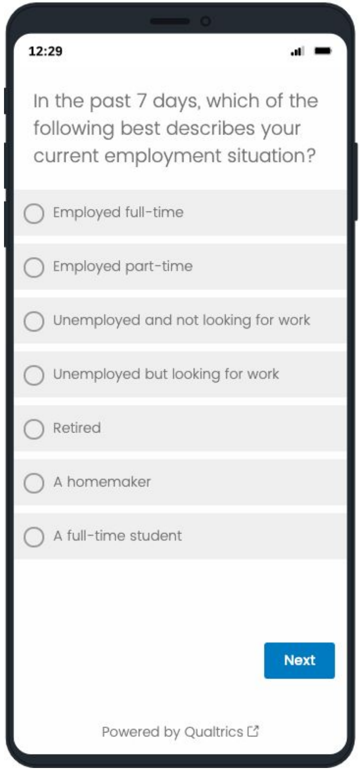
What is your gender?

○ Male

○ Female

○ Other

Demo: Employment



In the past 7 days, which of the following best describes your current employment situation?

○ Employed full-time

○ Employed part-time

○ Unemployed and not looking for work

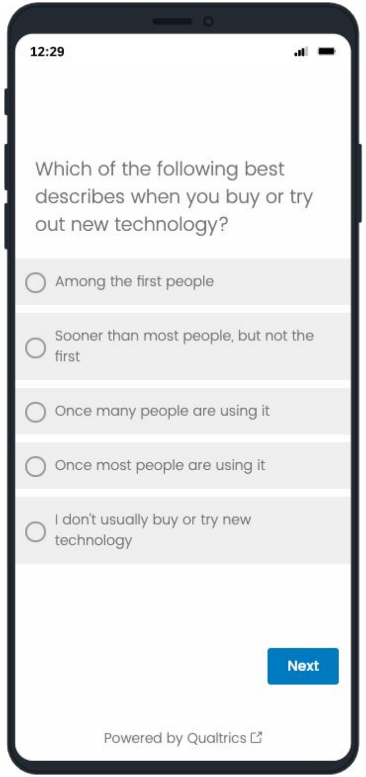
○ Unemployed but looking for work

○ Retired

○ A homemaker

○ A full-time student

Tech interest



Which of the following best describes when you buy or try out new technology?

○ Among the first people

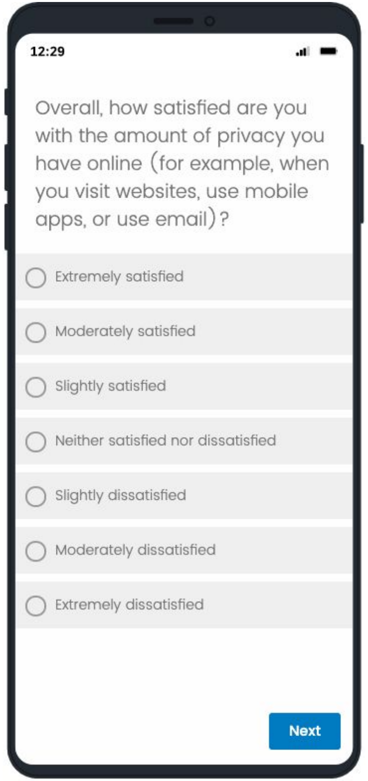
○ Sooner than most people, but not the first

○ Once many people are using it

○ Once most people are using it

○ I don't usually buy or try new technology

Privacy satisfaction



Overall, how satisfied are you with the amount of privacy you have online (for example, when you visit websites, use mobile apps, or use email)?

○ Extremely satisfied

○ Moderately satisfied

○ Slightly satisfied

○ Neither satisfied nor dissatisfied

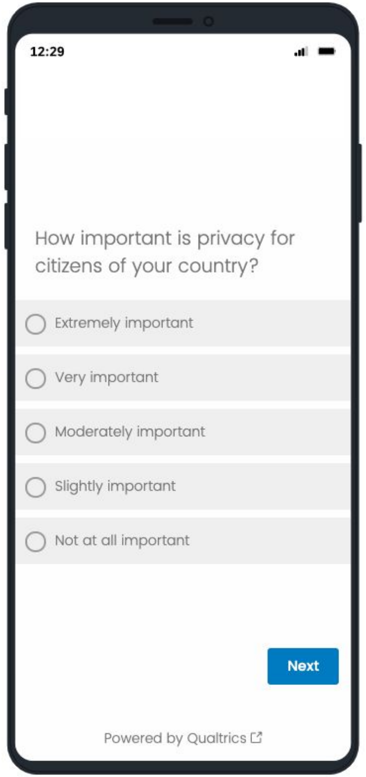
○ Slightly dissatisfied

○ Moderately dissatisfied

○ Extremely dissatisfied

Questions used in the surveys

Privacy importance



12:29

How important is privacy for citizens of your country?

○ Extremely important

○ Very important

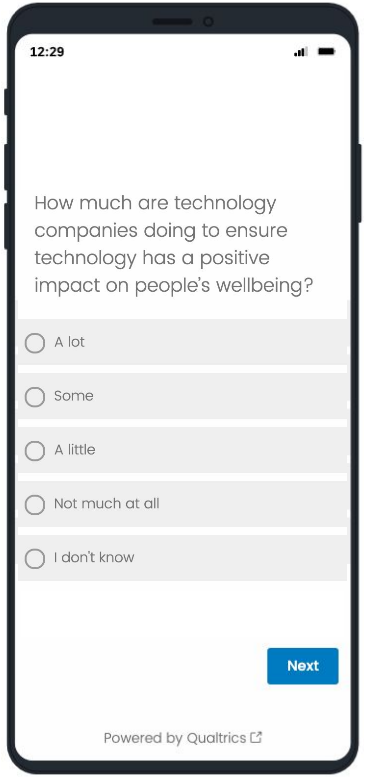
○ Moderately important

○ Slightly important

○ Not at all important

Powered by Qualtrics

Tech companies contribution



12:29

How much are technology companies doing to ensure technology has a positive impact on people's wellbeing?

○ A lot

○ Some

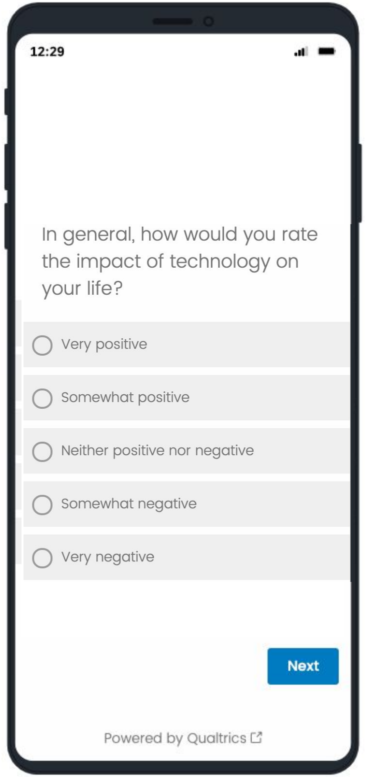
○ A little

○ Not much at all

○ I don't know

Powered by Qualtrics

Technology effects on life



12:29

In general, how would you rate the impact of technology on your life?

○ Very positive

○ Somewhat positive

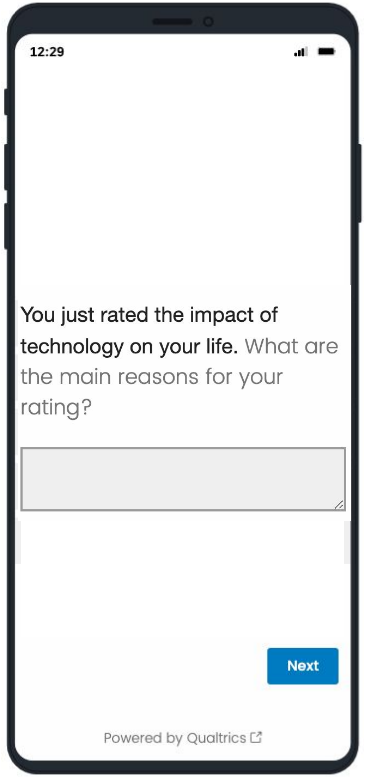
○ Neither positive nor negative

○ Somewhat negative

○ Very negative

Powered by Qualtrics

Open-ended question

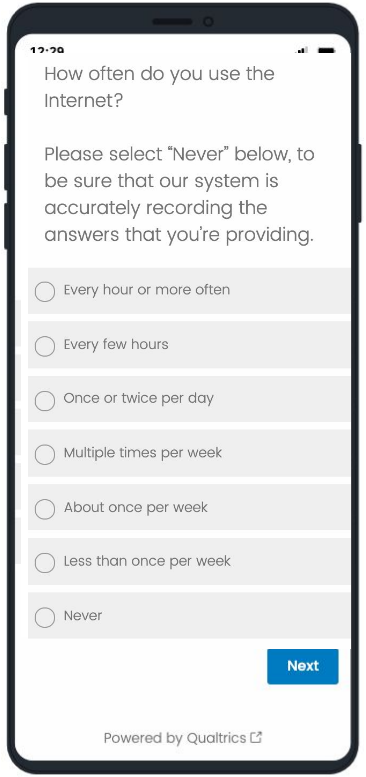


12:29

You just rated the impact of technology on your life. What are the main reasons for your rating?

Powered by Qualtrics

Attention checker



12:29

How often do you use the Internet?

Please select "Never" below, to be sure that our system is accurately recording the answers that you're providing.

○ Every hour or more often

○ Every few hours

○ Once or twice per day

○ Multiple times per week

○ About once per week

○ Less than once per week

○ Never

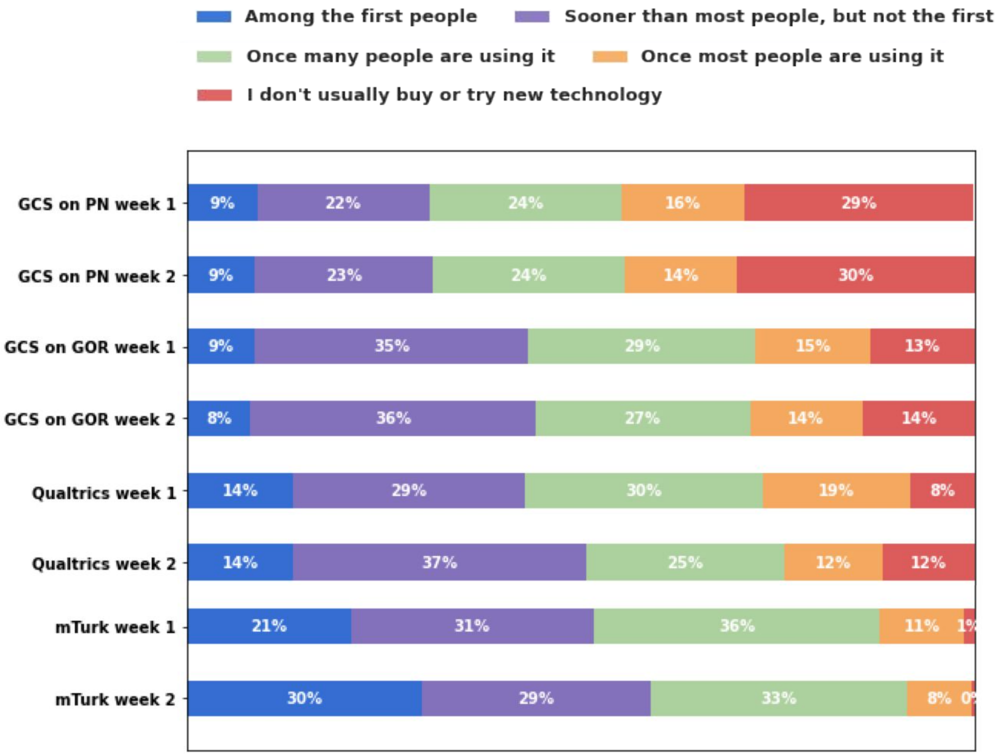
Powered by Qualtrics

Consistency metric

Sum of differences:
Absolute value of the difference (in percentage points) between each response option of the wave 1 data and the data measured in wave 2

In general, how would you rate the impact of technology on your life?	**Wave 1 value**	**Wave 2 value**	**Absolute difference**
Very positive	42	43	1
Somewhat positive	29	30	1
Neither positive nor negative	15	13	2
Somewhat negative	5	4	1
Very negative	10	9	1
		Sum of differences	**6**

Tech interest: Which of the following best describes when you buy or try out new technology?



Sum of differences:

3

5

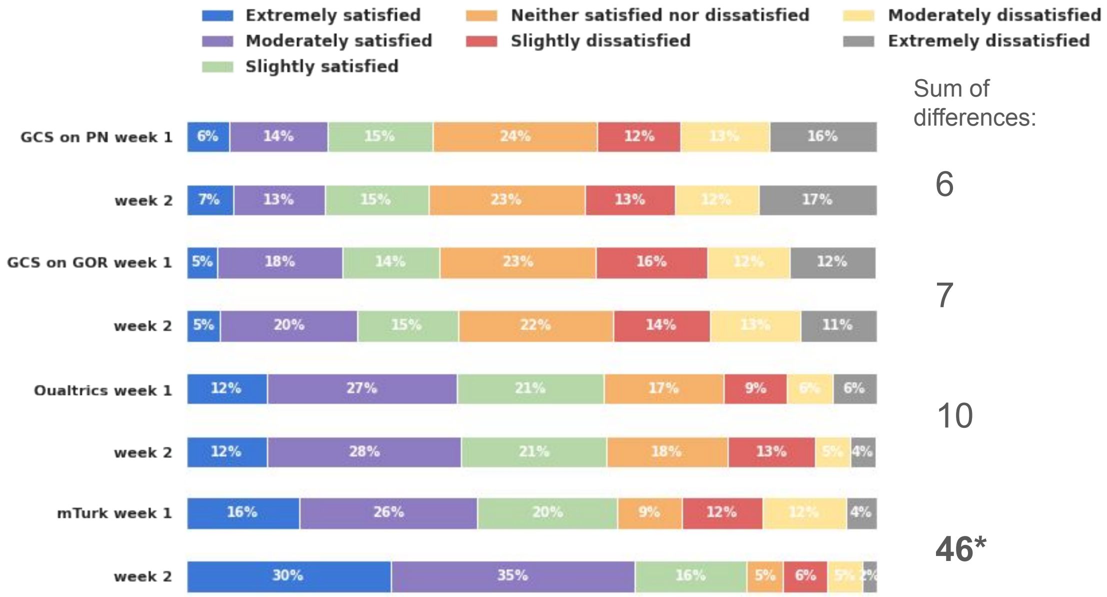
23*

18*

Qualtrics and MTurk have significantly different results

* Statistically significant at 0.01 using a weighted Welch test

Privacy satisfaction: Overall, how satisfied are you with the amount of privacy you have online?



Sum of differences:

6

7

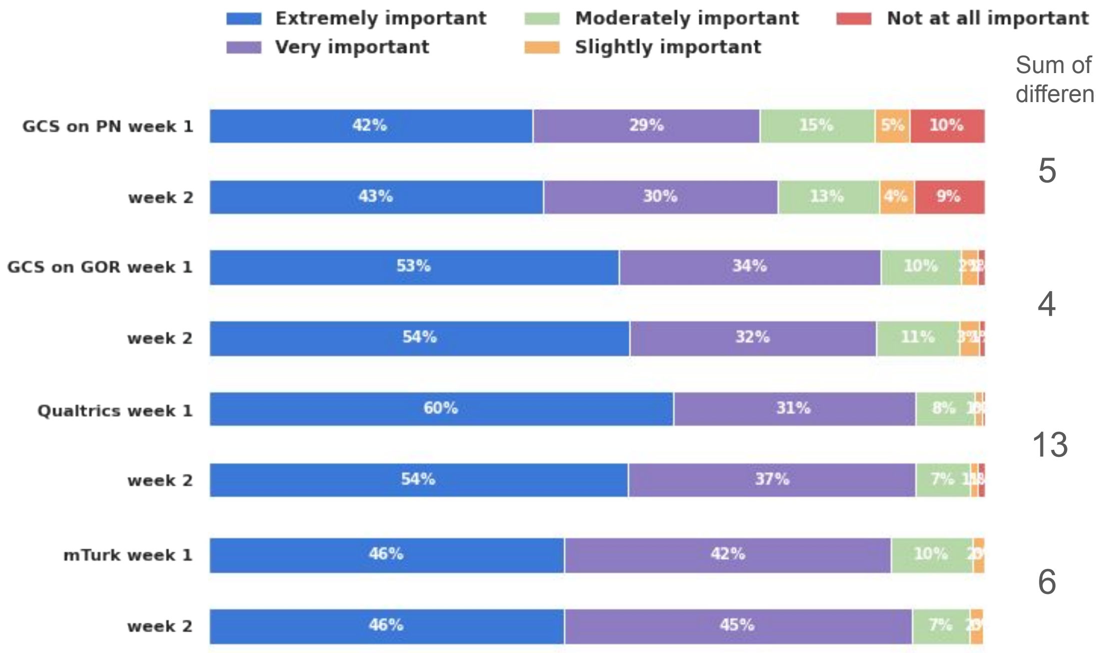
10

46*

mTurk is the only panel that shows a large and stat. sig. difference between the two runs

* Statistically significant at 0.01 using a weighted Welch test

Privacy importance: How important is privacy for citizens of your country?



Sum of differences:

5

4

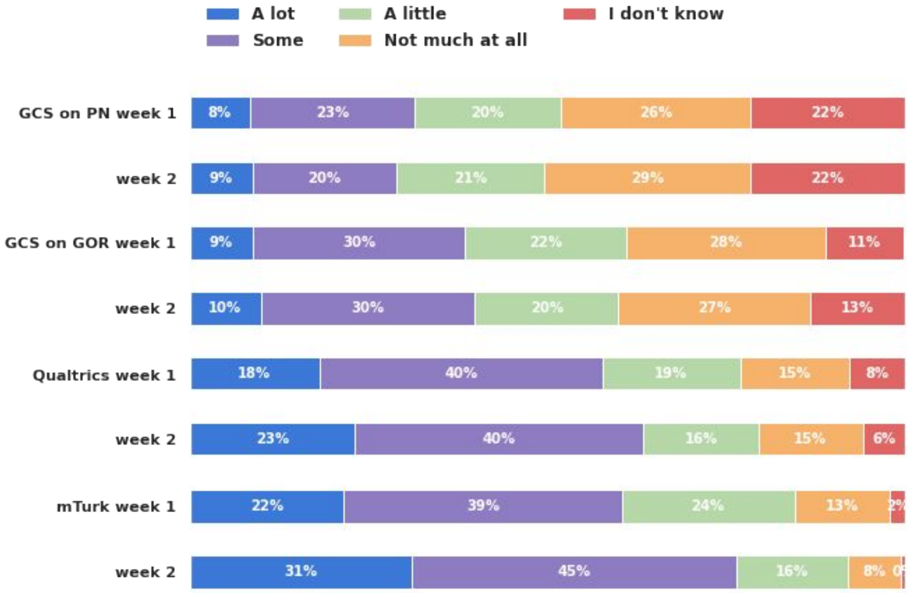
13

6

For all panels there were no significant differences between the two runs

* Statistically significant at 0.01 using a weighted Welch test

Tech companies contribution: How much are technology companies doing to ensure technology has a positive impact on people's wellbeing?



Sum of differences:

6

7

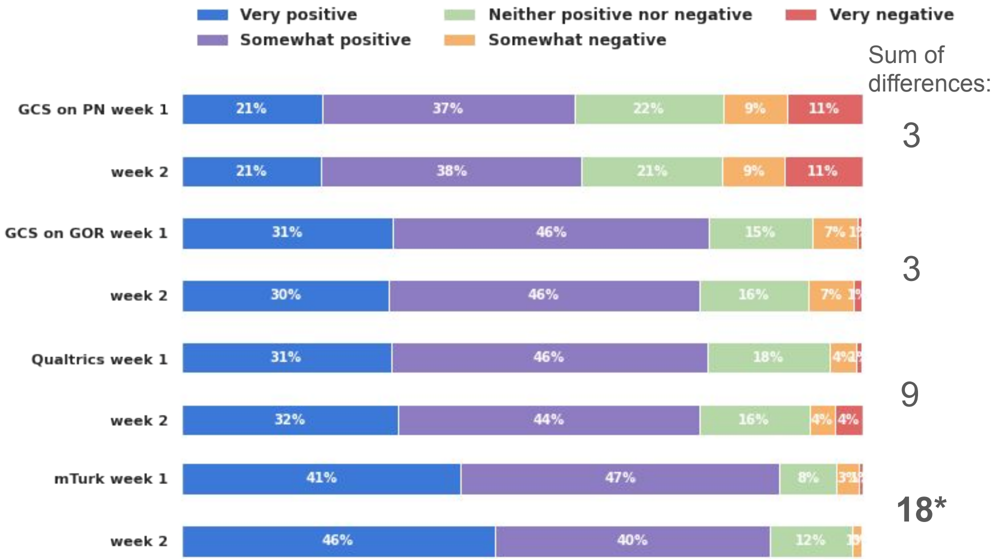
11*

32*

mTurk shows a large and stat. sig. difference between the two runs

* Statistically significant at 0.01 using a weighted Welch test

Technology effects on life: In general, how would you rate the impact of technology on your life?



Sum of differences:

3

3

9

18*

* Statistically significant at 0.01 using a weighted Welch test

mTurk is the only panel that shows a moderate stat. sig. difference between the two runs

Summary: mTurk shows the lowest consistency with 4 / 5 questions significantly different from wave 1 to wave 2 GCS on GOR and on PN shows the highest consistency

	GCS on PN	GCS on GOR	Qualtrics	mTurk
Tech interest	3	5	**23***	**18***
Privacy satisfaction	6	7	10	**46***
Privacy importance	5	4	13	6
Tech companies contribution	6	7	**11***	**32***
Technology effects on life	3	3	9	**18***
Average (5 attitude questions)	4.6	5.2	13.2	24.0

The numbers in the table are summed differences between week 1 and week 2 data.

Conclusions

Survey trackers are set up to measure changes over time

- Our study shows how 2 of the 4 panels obtained very unstable estimates in two measures just 2 weeks apart

- Amazon mTurk should not be considered a panel (it is not) and should not be used to track sentiment over time

Methods details

Age: Which age group best describes you?

	18-24	25-34	35-44	45-54	55-64	65+	Sum of differences
GCS on PN Week 1	8%	12%	20%	22%	21%	17%	
Week 2	8%	13%	17%	20%	23%	19%	**10**
GCS on GOR Week 1	8%	12%	20%	22%	21%	17%	
Week 2	15%	23%	21%	16%	13%	13%	**38**
Qualtrics Week 1	13%	19%	16%	12%	20%	20%	
Week 2	14%	21%	16%	12%	20%	18%	**7**
mTurk Week 1	6%	40%	28%	15%	8%	4%	
Week 2	3%	43%	33%	14%	5%	1%	**16**

Unweighted data and no statistical significance testing

Gender: What is your gender?

	Female	Male	Sum of differences
GCS on PN Week 1	52%	48%	
Week 2	52%	48%	**0**
GCS on GOR Week 1	43%	57%	
Week 2	46%	54%	**6**
Qualtrics Week 1	49%	51%	
Week 2	50%	50%	**2**
mTurk Week 1	40%	60%	
Week 2	43%	57%	**6**

Unweighted data and no statistical significance testing

Price and other technical characteristics

	GCS on PN	GCS on GOR	Qualtrics	mTurk
Price per respondent ranges*	$1-2	$1-2	$7-10	$1-2
Survey completion rate	40%	100%	89%	90%
Setup time	Few hours	Few hours	1-2 weeks	Few hours
Data Collection time*	33 hours	17 hours	31 hours	4 hours
Question limit	10	10	None	None
Question formats	Limited	Limited	Flexible	Flexible

* Specific to this set of surveys.
Different number and types of questions, as well as number of screeners will affect both price and collection time.

Questionnaire, sample size & data cleaning

Data collection weeks:

Week 1: Last week on June 2021
Week 2: Second week of July 2021

Questionnaire

The Qualtrics questionnaire is identical to Google survey with one question per page
Questionnaire in Google Surveys shown on these slides

Sample size

The sample size for each of the panels was of 1,500, to have enough statistical power

Data Cleaning

Google Surveys & Google Opinion Rewards have checks in place to remove invalid responses. See this help page here

Qualtrics provided a data cleaning service where according to their own proprietary algorithm, they removed 46 responses in wave 1 and 42 in wave 2. See speaker notes for details. Qualtrics also enabled: RelevantID, Bot detection, and Prevent multiple submissions for the study to exclude any potential duplicate or possibly fraudulent respondents.

mTurk does not provide any data cleaning processing

References

There are lots of studies on this topic but they all focus on accuracy. We list two review papers and some new studies

Review papers:

Callegaro, M., Villar, A., Yeager, D. S., & Krosnick, J. A. (2014). A critical review of studies investigating the quality of data obtained with online panels. In M. Callegaro, R. P. Baker, J. Bethlehem, A. S. Göritz, J. A. Krosnick, & P. J. Lavrakas (Eds.), *Online panel research. A data quality perspective* (pp. 23–53). Wiley. Link to PDF

Cornesse, C., Blom, A. G., Dutwin, D., Krosnick, J. A., De Leeuw, E. D., Legleye, S., Pasek, J., Pennay, D., Phillips, B., Sakshaug, J. W., Struminskaya, B., & Wenz, A. (2020). A review of conceptual approaches and empirical evidence on probability and nonprobability sample survey research. *Journal of Survey Statistics and Methodology*, *8*(1), 4–36. Link to PDF

New studies

Amaya, A., & Lau, A. (2021, June 9). *Measuring the risks of panel conditioning in survey research. Conditioning does not contribute significant error to panel estimates*. Link to PDF

Maru Blue (2020) Can we count on you? White Paper & Video