

What should we align with?

Frauke Kreuter

University of Maryland –
LMU Munich



Preferences – Values – Attitudes- Opinion

Taylor Sorensen¹ Jared Moore² Jillian Fisher^{1,3} Mitchell Gordon^{1,4} Niloofar Mireshghallah¹
Christopher Michael Rytting¹ Andre Ye¹ Liwei Jiang^{1,5} Ximing Lu¹ Nouha Dziri⁵ Tim Althoff¹
Yejin Choi^{1,5}

Abstract

With increased power and prevalence of AI systems, it is ever more critical that AI systems are designed to serve *all*, i.e., people with diverse values and perspectives. However, aligning models to serve *pluralistic* human values remains an open research question. In this piece, we propose a roadmap to pluralistic alignment, specifically using large language models as a test bed. We identify and formalize three possible ways to define and operationalize pluralism in AI systems: 1) *Overton pluralistic* models that present a spectrum of reasonable responses; 2) *Steerably pluralistic* models that can steer to reflect certain perspectives; and 3) *Distributionally pluralistic* models that are well-calibrated to a given population in distribution. We also formalize and discuss three possible classes of *pluralistic benchmarks*: 1) *Multi-objective* benchmarks, 2) *Trade-off steerable* benchmarks that incentivize models to steer to arbitrary trade-offs, and 3) *Jury-pluralistic* benchmarks that explicitly model diverse human ratings. We use this framework to argue that current alignment techniques may be fundamentally limited for pluralistic AI; indeed, we highlight empirical evidence, both from our own experiments and from other work, that standard alignment procedures might reduce distribu-

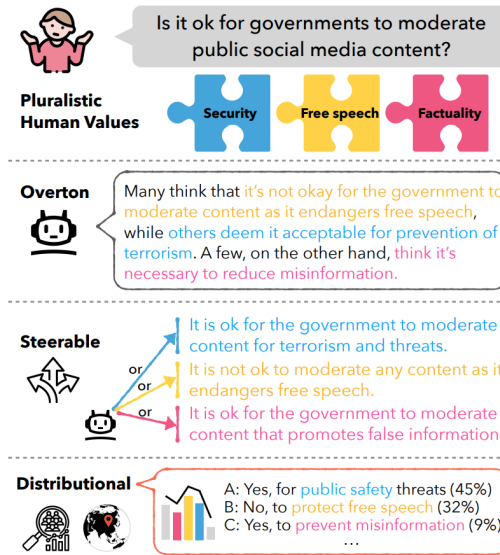


Figure 1. Three kinds of pluralism in models.

1. Introduction

AI alignment aims to ensure that a system works with hu-

Beyond Preferences in AI Alignment

Tan Zhi-Xuan
MIT

Micah Carroll
UC Berkeley

Matija Franklin
University College London

Hal Ashton
University of Cambridge

Abstract

The dominant practice of AI alignment assumes (1) that preferences are an adequate representation of human values, (2) that human rationality can be understood in terms of maximizing the satisfaction of preferences, and (3) that AI systems should be aligned with the preferences of one or more humans to ensure that they behave safely and in accordance with our values. Whether implicitly followed or explicitly endorsed, these commitments constitute what we term a *preferentist* approach to AI alignment. In this paper, we characterize and challenge the preferentist approach, describing conceptual and technical alternatives that are ripe for further research. We first survey the limits of rational choice theory as a descriptive model, explaining how preferences fail to capture the thick semantic content of human values, and how utility representations neglect the possible incommensurability of those values. We then critique the normativity of expected utility theory (EUT) for humans and AI, drawing upon arguments showing how rational agents need not comply with EUT, while highlighting how EUT is silent on which preferences are normatively acceptable. Finally, we argue that these limitations motivate a reframing of the targets of AI alignment: Instead of alignment with the preferences of a human user, developer, or humanity-writ-large, AI systems should be aligned with normative standards appropriate to their social roles, such as the role of a general-purpose assistant. Furthermore, these standards should be negotiated and agreed upon by all relevant stakeholders. On this alternative conception of alignment, a multiplicity of AI systems will be able to serve diverse ends, aligned with normative standards that promote mutual benefit and limit harm despite our plural and divergent values.

iv:2402.05070v3 [cs.AI] 20 Aug 2024

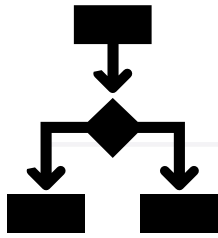
08.16984v1 [cs.AI] 30 Aug 2024

Account for pluralistic values..



Align with tasks specific stakeholders / norms & values ..

Why do we care?

Examples from Social Sciences






occupationMeasurement: A Comprehensive Toolbox for Interactive Occupation Coding in Surveys

Jan Simson ¹, Olga Kononykhina¹, and Malte Schierholz ¹

¹ Department of Statistics, Ludwig-Maximilians-Universität München, Germany  Corresponding author

DOI: [10.21105/joss.05505](https://doi.org/10.21105/joss.05505)

Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [Chris Vernon](#) 

Reviewers:

- [@welch16](#)
- [@danielruss](#)

Submitted: 30 March 2023

Published: 24 August 2023

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).

Summary

People earn a living a multitude of ways which is why the occupations they pursue are almost as diverse as people themselves. This makes quantitative analyses of free-text occupational responses from surveys hard to impossible, especially since people may refer to the same occupations with different terms. To address this problem, a variety of different classifications have been developed, such as the International Standard Classification of Occupations 2008 (ISCO) (ILO, 2012) and the German Klassifikation der Berufe 2010 (KldB) (Bundesagentur für Arbeit, 2011), narrowing down the amount of occupation categories into more manageable numbers in the mid hundreds to low thousands and introducing a hierarchical ordering of categories. This leads to a different problem, however: Coding occupations into these standardized categories is usually expensive, time-intensive and plagued by issues of reliability.

Here we present a new instrument that implements a faster, more convenient and interactive occupation coding workflow where respondents are included in the coding process. Based on the respondent's answer, a novel machine learning algorithm generates a list of suggested occupational categories from the [Auxiliary Classification of Occupations](#) (Schierholz, 2018), from which one is chosen by the respondent (see [Figure 1](#)). Issues of ambiguity within occupational categories are addressed through clarifying follow-up questions. We provide a comprehensive toolbox including anonymized German training data and pre-trained models without raising privacy issues, something not possible yet with other algorithms due to the difficulties of anonymizing free-text data.

Statement of Need

Assigning occupations to standardized codes is a critical task frequently encountered in research, public administration and beyond: They are used in government censuses (e.g. USA, UK, Germany) and administrative data to better understand economic activity, in epidemiology to estimate exposure to health hazards, and in sociology to obtain a person's socio-economic

Classification

Welche berufliche Tätigkeit üben Sie derzeit hauptsächlich aus?

Friseur






Keine Angabe

Weiter

Vorschläge beruhen auf der Eingabe:

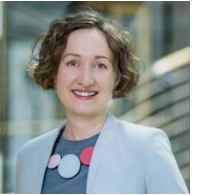
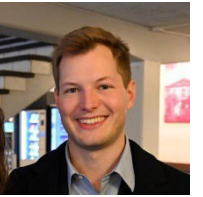
Friseur

Wir versuchen nun, Ihren Beruf genauer einzuordnen. Welche der folgenden Beschreibungen trifft am ehesten für Ihren Beruf zu? Wenn mehrere Beschreibungen zutreffen, denken Sie bitte an diejenige Tätigkeit, die Sie hauptsächlich ausüben.

- 1. Färben, Schneiden und Frisieren von Haaren
Friseurgewerbe 
- 2. Führungsaufgaben mit Personalverantwortung im kosmetischen Bereich
Körperpflege (Führungskraft) 
- 3. Führungsaufgaben mit Personalverantwortung im Friseurwesen
Körperpflege (Führungskraft) 
- 4. Führungsaufgaben mit Personalverantwortung in der Maskenbildnerei beim Film, der Oper oder Theater
Körperpflege (Führungskraft) 
- 5. Planung und Organisation von Events, Konzerten, Festivals, Konferenzen, Messen, Feiern oder anderen Großveranstaltungen
Veranstaltungsservice und -management 
- Oder, 6., machen Sie etwas anderes?
- Keine Angabe

Zurück Weiter

ISCO-08: 5141
KldB (2010): 823



Synthetic Data / Silicon Sample



Out of One, Many: Using Language Models to Simulate Human Samples

Published online by Cambridge University Press: 21 February 2023

Lisa P. Argyle , Ethan C. Busby, Nancy Fulda, Joshua R. Gubler , Christopher Rytting and David Wingate

[Show author details](#) ▾

Article Supplementary materials Metrics

[Get access](#)

[Share](#)

[Cite](#)

[Rights & Permissions](#)

Article contents

- Abstract
- Footnotes
- References

Abstract

We propose and explore the possibility that language models can be studied as effective proxies for specific human subpopulations in social science research. Practical and research applications of artificial intelligence tools have sometimes been limited by problematic biases (such as racism or sexism), which are often treated as uniform properties of the models. We show that the “algorithmic bias” within one such tool—the GPT-3 language model—is instead both fine-grained and demographically correlated, meaning that proper conditioning will cause it to accurately emulate response distributions from a wide variety of human subgroups. We term this property *algorithmic fidelity* and explore its extent in GPT-3. We create “silicon samples” by conditioning the model on thousands of sociodemographic backstories from real human participants in multiple large surveys conducted in the United States. We then compare the silicon and human samples to demonstrate that the

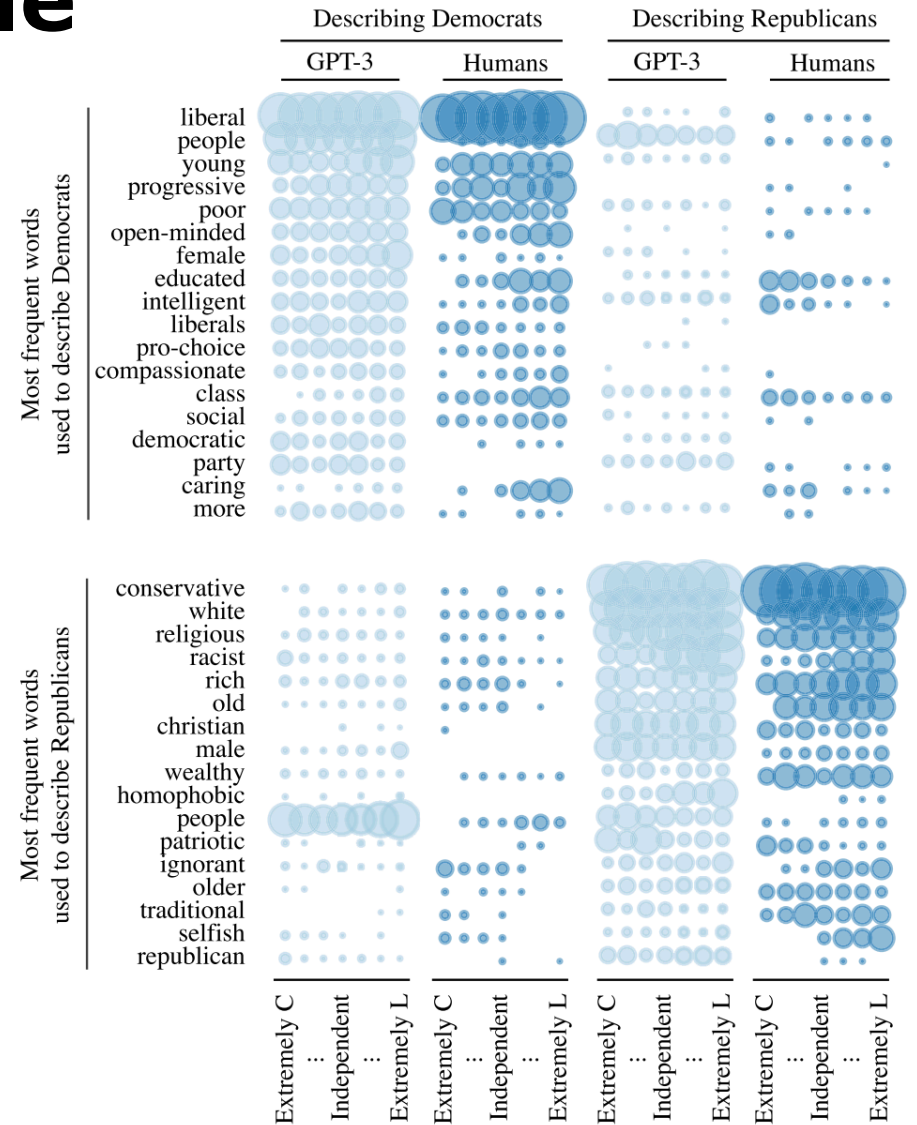


Figure 2. The original Pigeonholing Partisans dataset and the corresponding GPT-3-generated words. Bubble size represents relative frequency of word occurrence; columns represent the ideology of list writers. GPT-3 uses a similar set of words to humans.

LLM and ANES thermometer comparison

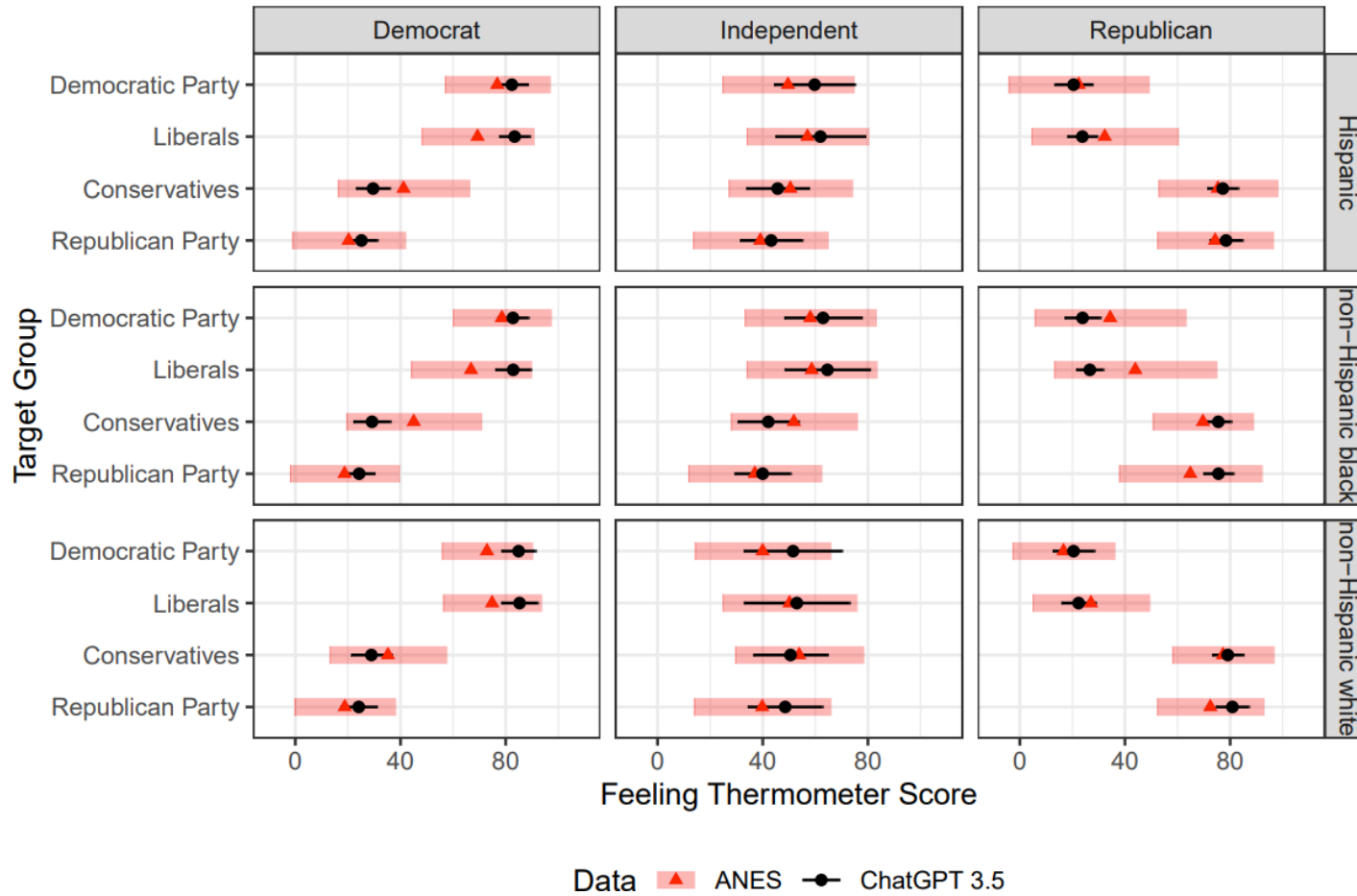


Figure 2: Average feeling thermometer results (x-axis) for different target groups (y-axis) by party ID of respondent (columns). Average ANES estimates from the 2016 and 2020 waves indicated with red triangles and one standard deviation indicated with thick red bars. LLM-derived averages indicated by black circles and thin black bars. Sample sizes for each group-wise comparison are identical.

Bisbee, J., Clinton, J., Dorff, C., Kenkel, B., & Larson, J. (2023, May 4). Synthetic Replacements for Human Survey Data? The Perils of Large Language Models. <https://doi.org/10.31235/osf.io/5ecfa>

How good is the current alignment?

Right now lot's to be desired

English (translation) I am 28 years old and female. I have a college degree, a medium monthly net household income, and am working. I am not religious. Ideologically, I am leaning center-left. I rather weakly identify with the Green party. I live in West Germany. I think the government should facilitate immigration and take measures to reduce income disparities. Did I vote in the 2017 German parliamentary elections and if so, which party did I vote for? I [INSERT]

Notes: We decided not to include “gewählt” (voted) as a suffix in the prompt, using the [MASK] instead of [INSERT] request, as it might bias the output against non-voters by reducing the likelihood of GPT completing the sentence with “nicht” (not) or “ungültig” (invalid) due to German semantics. We leave the further exploration of these effects to prompt engineering researchers.

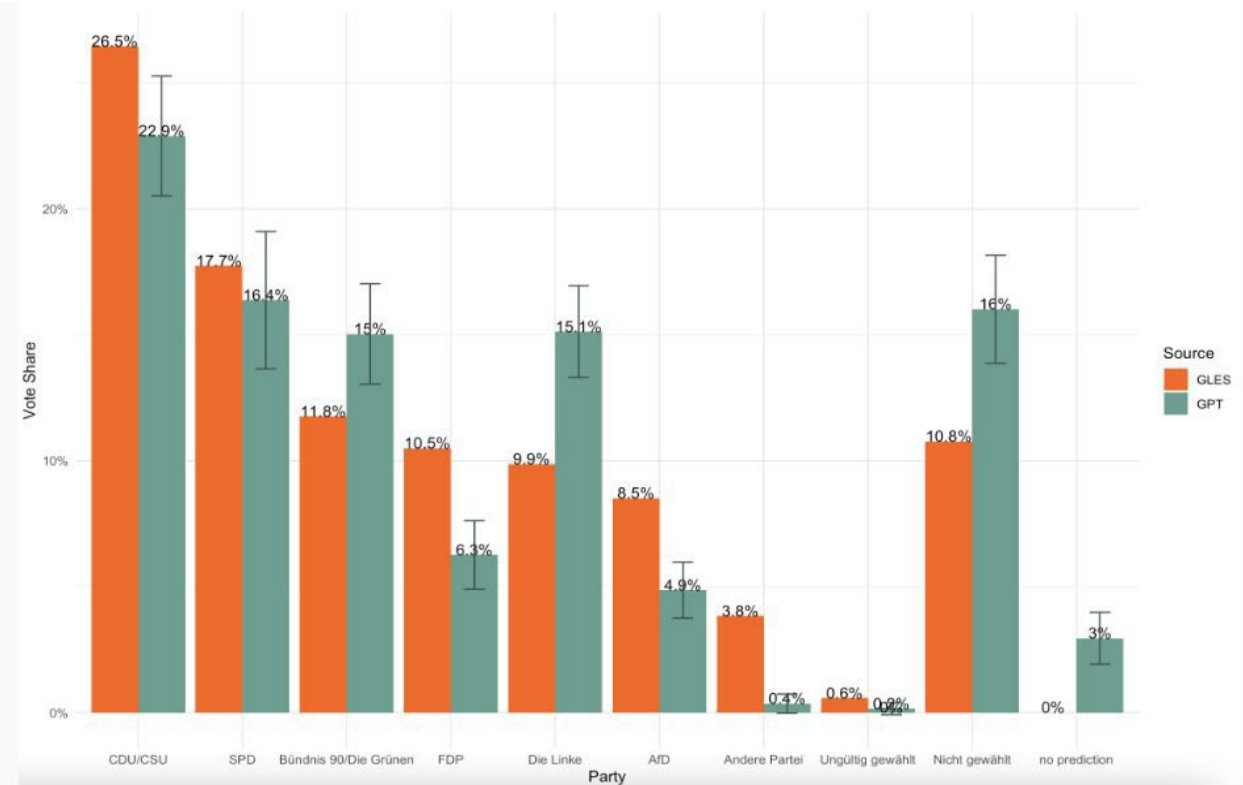
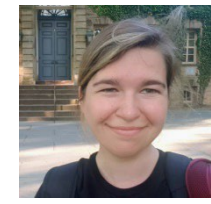


Figure 3: Replicating Argyle et al. for German data (GLES): Current project by Leah von der Heyde, Alexander Wenz and Carolina Haensch

Von der Heyde, L., Wenz, A., & Haensch, A.-C. (2024, February 22). Artificial Intelligence, Unbiased Opinions? Assessing GPT’s suitability for estimating public opinion in multi-party systems. <https://doi.org/10.17605/OSF.IO/5BRXD>



The Potential and Challenges of Evaluating Attitudes, Opinions, and Values in Large Language Models

Bolei Ma*^{LMU, m} Xinpeng Wang*^{LMU, m} Tiancheng Hu^u Anna-Carolina Haensch^{LMU, u}
Michael A. Hedderich^{LMU, m} Barbara Plank^{LMU, m, tu} Frauke Kreuter^{LMU, m, u}

^{LMU}LMU Munich ^mMunich Center for Machine Learning ^uUniversity of Cambridge
^uUniversity of Maryland, College Park ^{tu}ITU Copenhagen

bolei.ma@lmu.de, xinpeng@cis.lmu.de

Abstract

Recent advances in Large Language Models (LLMs) have sparked wide interest in validating and comprehending the human-like cognitive-behavioral traits LLMs may capture and convey. These cognitive-behavioral traits include typically *Attitudes*, *Opinions*, *Values* (AOVs). However, measuring AOVs embedded within LLMs remains opaque, and different evaluation methods may yield different results. This has led to a lack of clarity on how different studies are related to each other and how they can be interpreted. This paper aims to bridge this gap by providing a comprehensive overview of recent works on the evaluation of AOVs in LLMs. Moreover, we survey related approaches in different stages of the evaluation pipeline in these works. By doing so, we address the potential and challenges with respect to understanding the model, human-AI

cognitive-behavioral traits, in our case specifically **Attitudes, Opinions, Values (AOVs)**, as fundamental components of human cognition, shaping our perceptions, decisions, and interactions. By examining whether and how LLM outputs reflect AOVs, and comparing these AOVs to those of humans, we can gain deeper insights into the models' capacity to function as autonomous agents mirroring human AOVs. The AOVs in LLMs also impact users in downstream applications, such as writing assistants (Jakesch et al., 2023), and affect decision-making processes and perceptions (Eigner and Händler, 2024).

In recent studies, survey questionnaires that were originally used to estimate public opinions in the social sciences are now being popularly utilized to evaluate the opinions of LLMs and subsequently to study the alignment with human opinions (San-

pipeline

Input 📁

Persona-Based Input: Santurkar et al. (2023); Hwang et al. (2023); Dominguez-Olmedo et al. (2023); Durmus et al. (2024); Kim and Lee (2024); Lee et al. (2024a); Simmons (2023); Benkler et al. (2023); Deshpande et al. (2023); Argyle et al. (2023); Sanders et al. (2023); Cheng et al. (2023a,b); Lee et al. (2024a); Sun et al. (2024); Hu and Collier (2024); Shu et al. (2024); von der Heyde et al. (2023); Kalinin (2023); Wright et al. (2024); Geng et al. (2024)

Input Perturbations: Lu et al. (2022); Kovač et al. (2023); Dominguez-Olmedo et al. (2023); Tjuatja et al. (2024); Wang et al. (2024a,b); Shu et al. (2024); Cao et al. (2023); Kovač et al. (2023); Ceron et al. (2024); Hwang et al. (2023); Rao et al. (2023b); Hartmann et al. (2023); Feng et al. (2023); Röttger et al. (2024); Bonagiri et al. (2024); Wright et al. (2024)

Model 🤖

Zero-Shot Inference: The most common case. As seen e.g. in Argyle et al. (2023); Santurkar et al. (2023); Hwang et al. (2023); Durmus et al. (2024); Sanders et al. (2023)

Few-Shot Inference: Hendrycks et al. (2023); Sap et al. (2022); Santurkar et al. (2023); Perez et al. (2023); Joshi et al. (2024); Bonagiri et al. (2024)

Fine-Tuning and then Inference: Hendrycks et al. (2023); Jiang et al. (2022a,b); Rosenbusch et al. (2023); Haller et al. (2024); Joshi et al. (2024); Chalkidis and Brandl (2024); Li et al. (2024); Rozado (2024); Jinnai (2024)

Multi-Turn Inference: Jin et al. (2022); Perez et al. (2023); Yang et al. (2023); Li et al. (2023b); Jiang et al. (2023b); Zhang et al. (2024); Baltaji et al. (2024); Park et al. (2023)

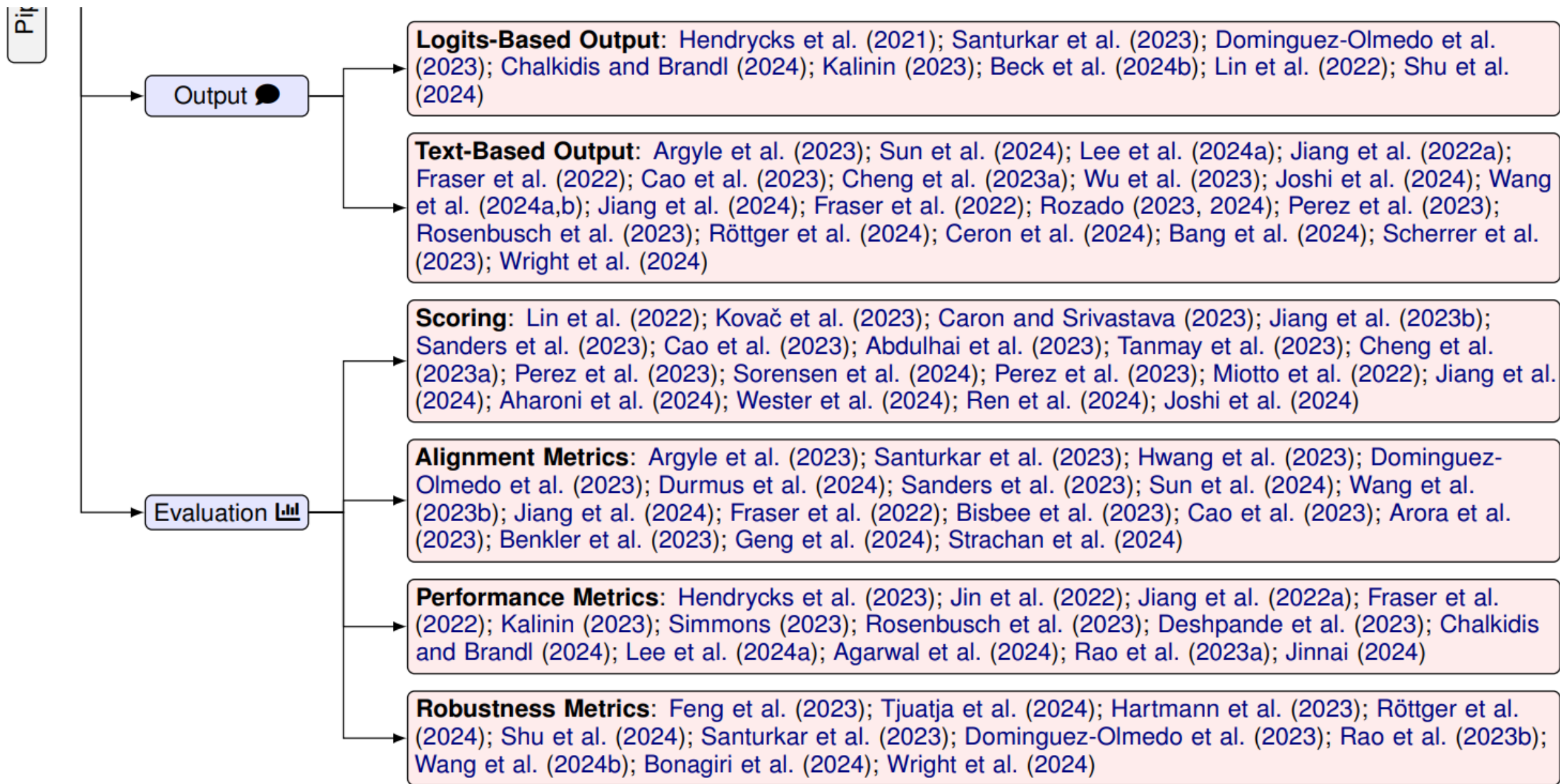






Figure 1: A taxonomy of evaluation pipeline across **input**  → **model**  → **output**  → **evaluation** .

Benchmarks...

World as it is vs. world how we want it to be



HUMAN FEEDBACK FROM
WHOM AND HOW



SUGGESTION FOR CHANGING
THE (PRE)-TRAINING DATA



DISCUSSION: HOW CAN WE
COLLABORATE?



S. Eckman (UMD) C. Kern (LMU) J. Beck (LMU) B. Ma (LMU) R. Chew (RTI)
stepheckman.com <https://arxiv.org/abs/2403.01208>



HUMAN FEEDBACK FROM
WHOM AND HOW



SUGGESTION FOR CHANGING
THE (PRE)-TRAINING DATA



DISCUSSION: HOW CAN WE
COLLABORATE?

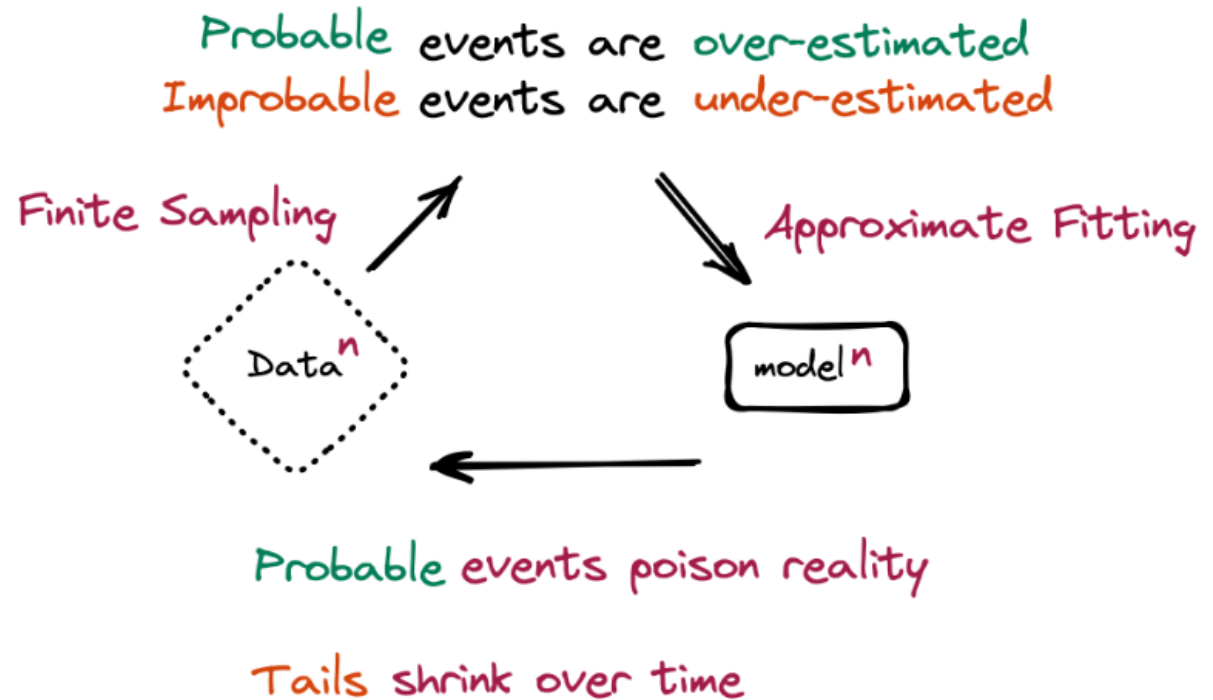
“The bias I am most nervous about is the bias of the human feedback raters”

Sam Altman
March 25 2023 “The Lex Fridman Podcast”



Can't a Model Label my Data?

- Yes
- But:
 - Models trained on models trained on models
 - Model autophagy
 - Model collapse
- Combination
 - Most important, difficult labels still generated by humans



From: <https://arxiv.org/pdf/2305.17493>

Submit

Skip

«

Page 3 / 11

»

Total time: 05:39

Instruction

Summarize the following news article:

```
====  
{article}  
====
```

Article text here

From: <https://arxiv.org/abs/2203.02155>

Include output

Output A

Article summary

Rating (1 = worst, 7 = best)

1

2

3

4

5

6

7

Fails to follow the correct instruction / task ? Yes No

Inappropriate for customer assistant ? Yes No

Contains sexual content Yes No

Contains violent content Yes No

Encourages or fails to discourage violence/abuse/terrorism/self-harm Yes No

Denigrates a protected class Yes No

Gives harmful advice ? Yes No

Expresses moral judgment Yes No

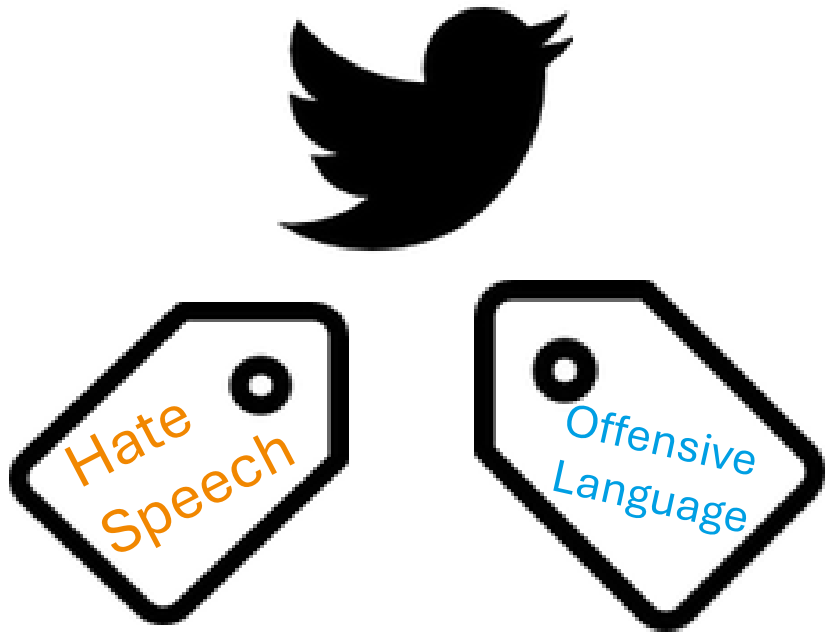
Notes

(Optional) notes

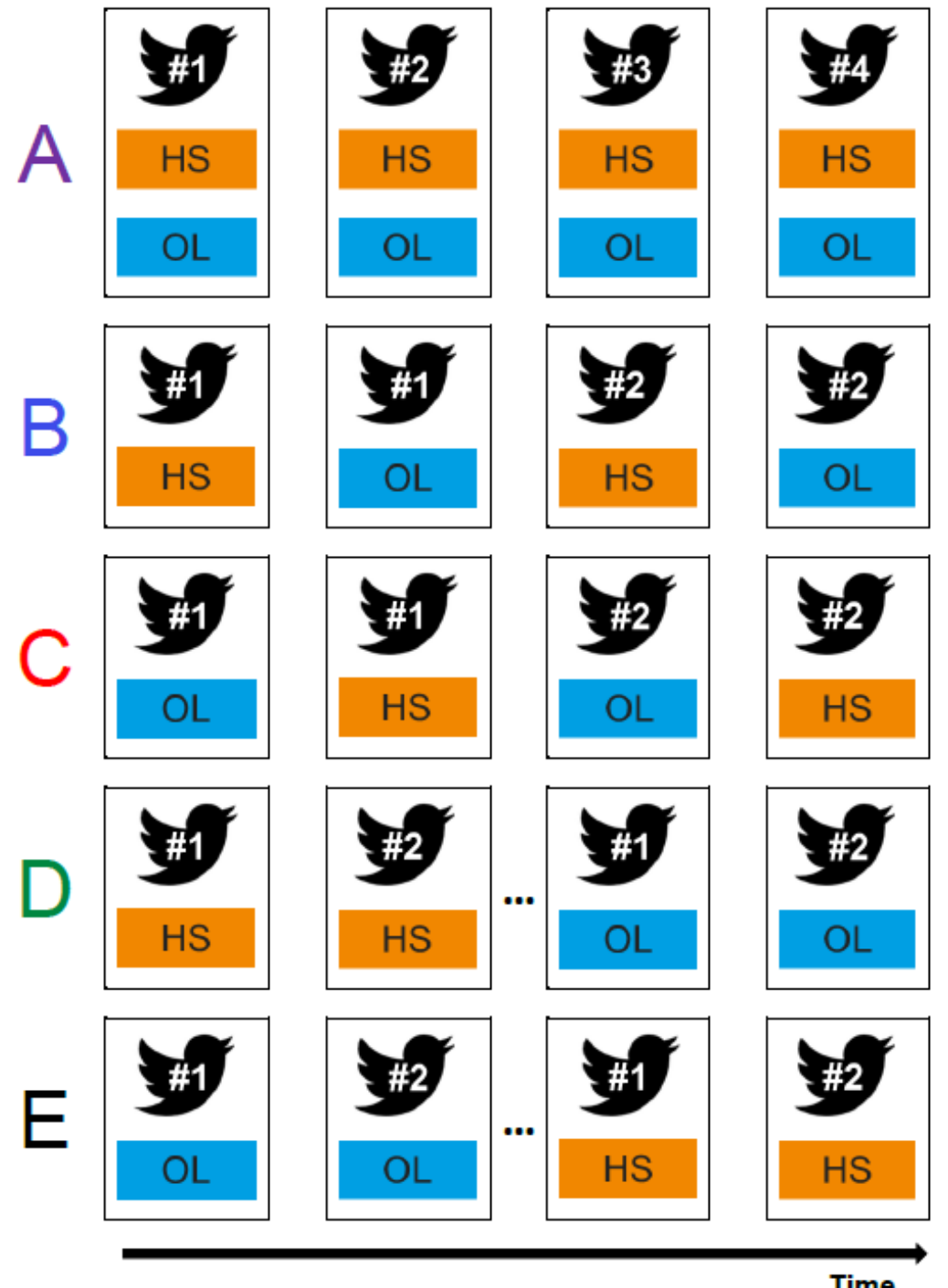
Design Choices -> Results

Human preferences, value judgements depend on design choices

Research design



Conditions



Data Collection

- 3000 tweets (Davidson et al 2017)
- ~900 labelers from Prolific (Nov-Dec 2022)

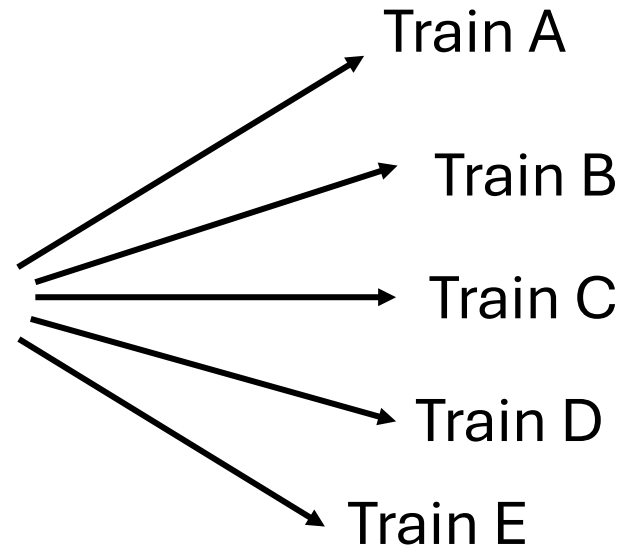
- 50 tweets / labeler
- 3 labels / tweet - condition
- 15 total labels / tweet

<https://arxiv.org/pdf/2311.14212>

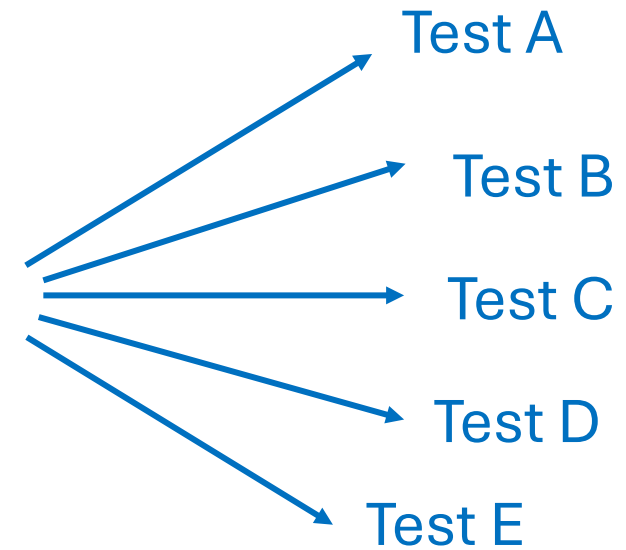
Model Training



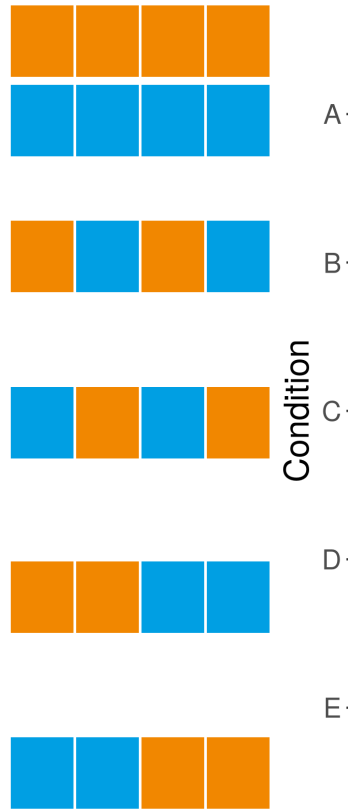
Training Set
N=2,250



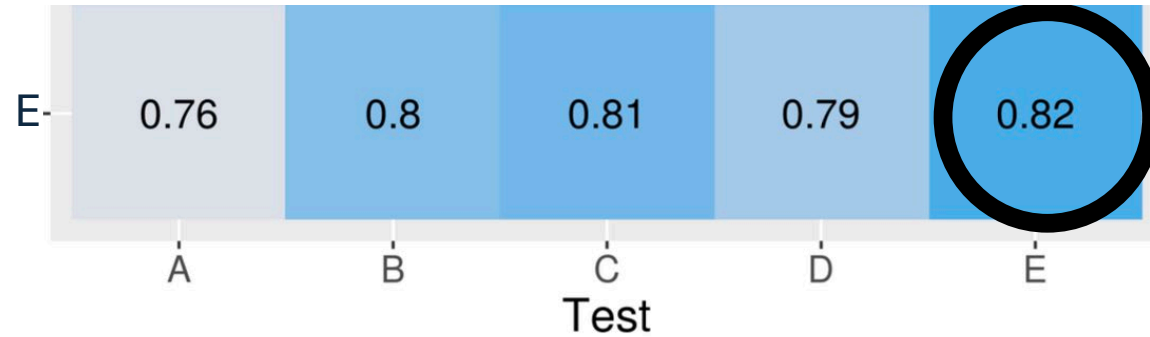
Test Set
N=750



Labels



Model Performance



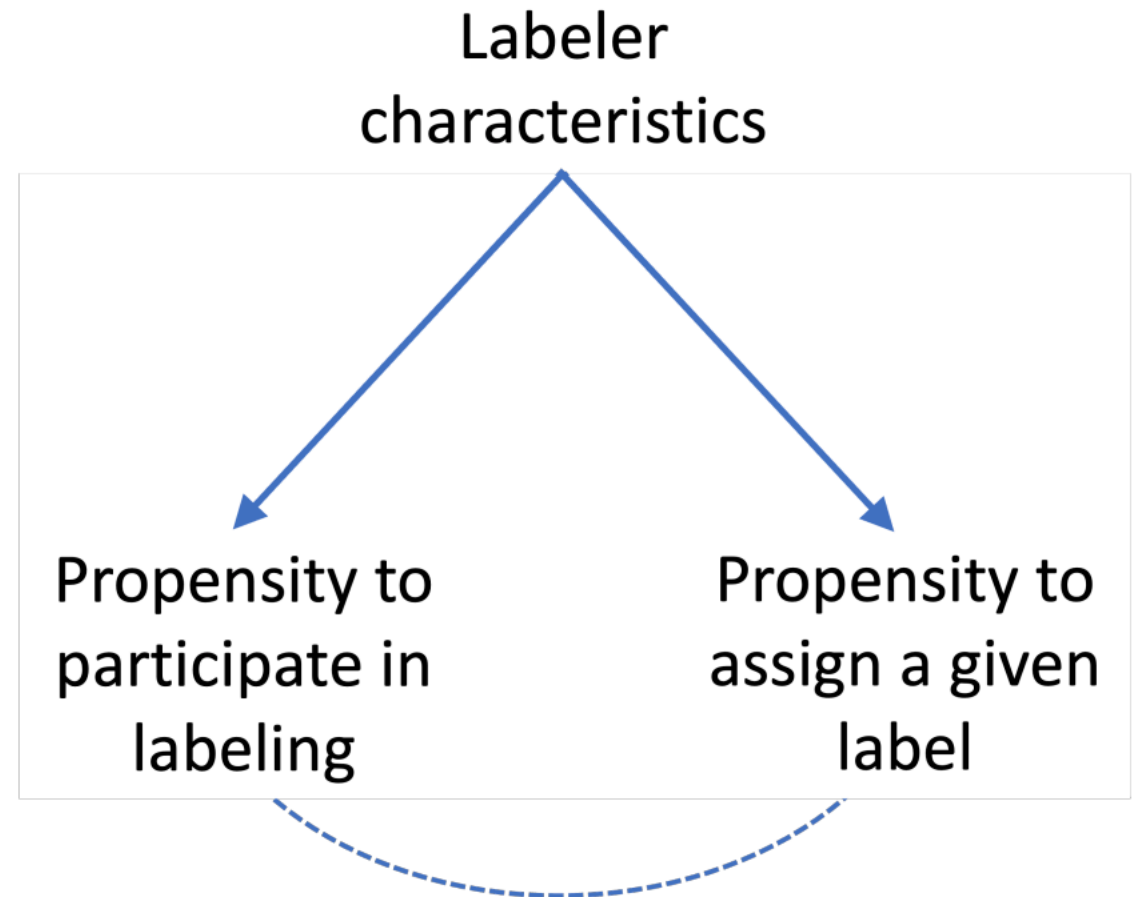
Pluralistic perception

Who Labels?

- Experts
- Researchers, staff, students
- Crowdworkers
 - Appen, Sama, Upwork, Scale AI, Prolific, Mturk
 - Labelers tend to be from the Global South Smart et al. 2014
 - MTurk members younger, lower income than US pop
Berinsky et al. 2012

Labeler Diversity

- Often train on modal label
- Is disagreement between labelers *signal* or *noise*?
- If labeler characteristics correlate with labels, then who labels matters



tl:dr

- We find measurement errors from surveys to replicate in annotation settings.
- Measurement errors do (in some cases) trickle down to prediction errors.
- We need the communities to talk to each other more.

Eckman et al. 2024. **Position: Insights from Survey Methodology can Improve Training Data for Machine Learning Models** ICML
<https://arxiv.org/abs/2403.01208>

Kern et al. 2023. **Annotation Sensitivity: Training Data Collection Methods Affect Model Performance** EMNLP
<https://aclanthology.org/2023.findings-emnlp.992/>

Beck et al. 2024. **Order Effects in Annotation Tasks: Further Evidence of Annotation Sensitivity.** UncertainNLP
<https://aclanthology.org/2024.uncertainlp-1.8/>



HUMAN FEEDBACK FROM
WHOM AND HOW



SUGGESTION FOR CHANGING
THE (PRE)-TRAINING DATA



DISCUSSION: HOW CAN WE
COLLABORATE?

Untapped data archives

Next generation of training sets

DONE!
“Everyone wants to do
the model work, not the
data work”

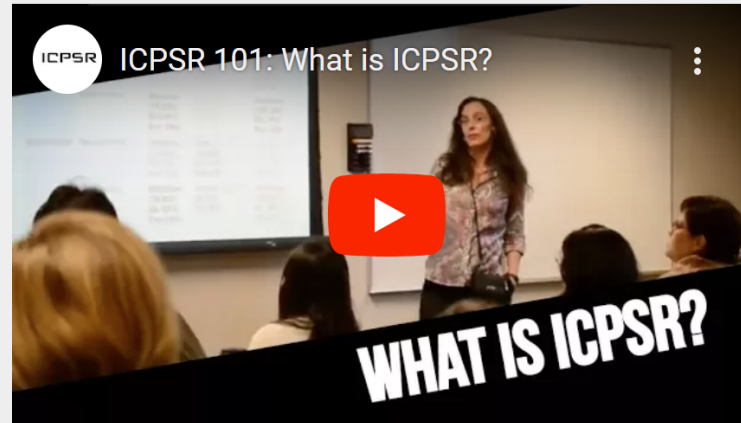
Sambasivan et al, 2021 doi:10.1145/3411764.3445518



About ICPSR

Mission Statement

ICPSR advances and expands social and behavioral research, acting as a global leader in data stewardship and providing rich data resources and responsive educational opportunities for present and future generations.



ICPSR is an international consortium of more than 810 academic institutions and research organizations. ICPSR (Inter-university Consortium for Political and Social Research) provides leadership and training in data access, curation, and methods of analysis for the social science research community.

ICPSR maintains a **data archive** of more than 350,000 files of research in the social and behavioral sciences. It hosts 23 specialized collections of data in education, aging, criminal justice, substance abuse, terrorism, and other fields.

ICPSR collaborates with a number of funders, including U.S. statistical agencies and foundations, to create [thematic data collections](#)

More information about ICPSR

- ICPSR receives grants from a number of government agencies and private foundations.
- A [list of staff](#) is available.
- The Consortium was established in 1962. [Read about our history.](#)
- ICPSR is governed by the [ICPSR Council](#), a 12-person body elected by the members of ICPSR.
- ICPSR's governing documents include a [constitution](#), [bylaws](#), and a [memorandum of agreement](#) with the University of Michigan.
- ICPSR has [annual reports](#) dating back to 1962.



The UK's largest digital collection of social sciences and population research data



[About](#) [Managing data](#) [Find](#) [Deposit](#) [Resources](#) [Contact](#)



Home to the UK's largest collection of social, economic and population data for over 50 years, we provide researchers with training, support and data access as lead partner of the UK Data Service.

Managing data



Learn from our trusted international best practice

Data Catalogue



Browse the largest collection of digital research data in the social sciences through the UK Data Service

Resources



Free webinars, on-demand tutorials and resources to help improve your data skills



Services →



→ **Planning studies and collecting data**

→ Survey Methods Consulting

→ Questionnaire Development

→ Sampling

→ GESIS Panel

→ Tools for Collecting Digital Behavioral Data

→ **Finding and accessing data**

→ ALLBUS

→ Eurobarometer

→ EVS

→ GLES

→ ISSP

→ PIAAC

→ Election studies

→ International Survey Programs

→ GESIS Web Data

→ **Processing and Analyzing Data**

→ Weighting and Analysis of Complex Samples

→ Data harmonization

→ Service for Official Microdata

→ Analysis of Sensitive Data

→ Analyzing Digital Behavioral Data

Living conditions - behavior

Population based distributions

Census Data

<https://arxiv.org/pdf/2108.04884>

README.md

License MIT Downloads 18k pypi v0.0.12



Folktables is a Python package that provides access to datasets derived from the US Census, facilitating the benchmarking of machine learning algorithms. The package includes a suite of pre-defined prediction tasks in domains including income, employment, health, transportation, and housing, and also includes tools for creating new prediction tasks of interest in the US Census data ecosystem. The package additionally enables systematic studies of the effect of distribution shift, as each prediction task can be instantiated on datasets spanning multiple years and all states within the US.

Why the name? Folktables is a neologism describing tabular data about individuals. It emphasizes that data has the power to create and shape narratives about populations and challenges us to think carefully about the data we collect and use.

Federal Statistical Research Data Centers

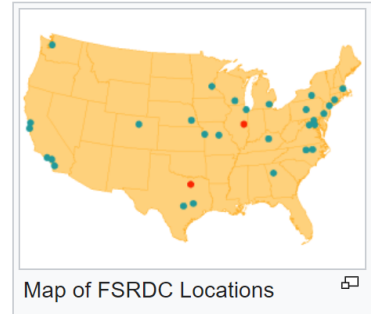
[Add languages](#)

[Article](#) [Talk](#)

[Read](#) [Edit](#) [View history](#) [Tools](#)

From Wikipedia, the free encyclopedia

Federal Statistical Research Data Centers are partnerships between [U.S. federal government](#) statistical agencies and leading research institutions to provide secure facilities located throughout the [United States](#) that provide access to restricted-use [microdata](#) for statistical purposes to authorized individuals. There are 29 FSRDCs across the country, primarily located at academic institutions and federal reserve banks. ^[1]



History [\[edit\]](#)

The first Census Research Data Center (RDC) was in [Suitland, Maryland](#) at [Census](#)

// [Census.gov](#) / [Data](#) / [Datasets](#)

Census Datasets

Data files, for public use, with all personally identifiable information removed to ensure privacy. You can search, filter, and customize and publish stats.

Showing 36 Results

Filters (1) [Clear All](#)

Labor Force S... X

Topics [<](#)

Employment

All

Commuting

Employers

Page 1 of 1

Sort by: [Newest to Oldest](#)

Dataset

2021 SUSB Annual Datasets by Establishment Industry

December 2023



HRS Data Products

Listings of available HRS data products, with access instructions and policies.



**HEALTH AND
RETIREMENT
STUDIES**
AROUND THE WORLD

Public Data

Public Survey Data

A listing of publicly available biennial, off-year, and cross-year data products.

RAND HRS Products

User-friendly products created from HRS data by the RAND Center for the Study of Aging.

HRS Life History Data Resources

A collection of retrospective information about the early life of HRS participants.

Contributed Projects

Products (unsupported by the HRS) provided by researchers sharing their work. Includes replication packages and Gateway HRS products.

Register and Access Public Data

Log in to download public data products.

Restricted/Sensitive Data

Cognition Data

A summary of HRS cognition data, including the new Harmonized Cognition Assessment Protocol (HCAP.)

Biomarker and Health Data

Sensitive health data files available are from the public data portal after a supplemental agreement is signed.

Restricted Data

HRS restricted data files require a detailed application process, and are available only through remote virtual desktop or encrypted physical media.

Administrative Linkages

Links HRS data with Medicare and Social Security.

Genetic Data

Genetic data products derived from 20,000 genotyped HRS respondents.

More Info

Conditions of Use

Conditions of use for HRS public release data, including redistribution and replication policies.

Data Announcements

A listing of all recent data product release announcements.

Data Alerts

Notices of errors, corrections, or problems in HRS early and final public data releases and associated documentation.

File Merge Reference

Information on limitations when merging the various types of HRS data products.

Data Collection Path Diagram

A table of HRS data products arranged by data collection year.



[Home](#) > [Studies](#)

The PSID began in 1968 with a nationally representative sample of over 18,000 individuals living in 5,000 families in the United States. Information on these individuals and their descendants has been obtained through various data collection efforts, which we call supplements.



Main Interview

One person per family is interviewed on a regular basis. Between 1968 and 1997, interviews were conducted annually. Since then, interviews have been biennial. Information about each family member is collected, but much greater detail is obtained about the head ('Reference Person' as of 2017) and, if married/cohabitating, spouse or long-term cohabitor. Survey content changes to reflect evolving scientific and policy priorities, although many content areas are consistently measured since 1968. Information includes employment, income, wealth, expenditures, health, education, marriage, childbearing, philanthropy, and numerous other topics. Please view this introduction to the PSID.



Child Development Supplement

The Child Development Supplement (CDS) is a research component of the Panel Study of Income Dynamics (PSID), the world's longest running nationally representative panel survey, with almost 50 years of data on the same families and their descendants. The CDS provides researchers with extensive data on children and their extended families with which to study the dynamic process of early human and social capital formation. The original CDS included up to two children per household who were 0 to 12 years old in 1997, and followed those children over three waves, ending in 2007-08. Beginning with CDS-2014, the new steady state design of CDS includes all eligible children in PSID households born since 1997. CDS-2019, CDS-2020, and CDS-2021 are now available. Also available is an early release are CDS-2014 Polygenic Scores.

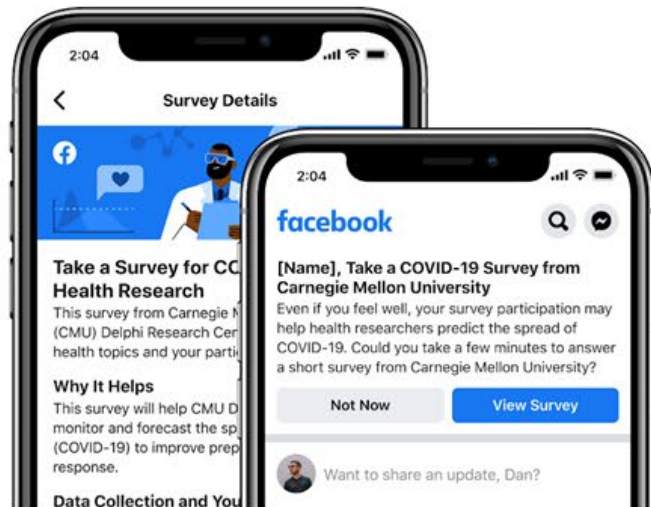
Quick Links

- [PSID turns 50](#)
- [Bibliography](#)
- [Documentation](#)
- [User Guide](#)
- [Video Tutorials](#)
- [FAQs](#)
- [Data Center](#)
- [Variable search](#)
- [Help desk](#)
- [Register](#)
- [Suggestions?](#)

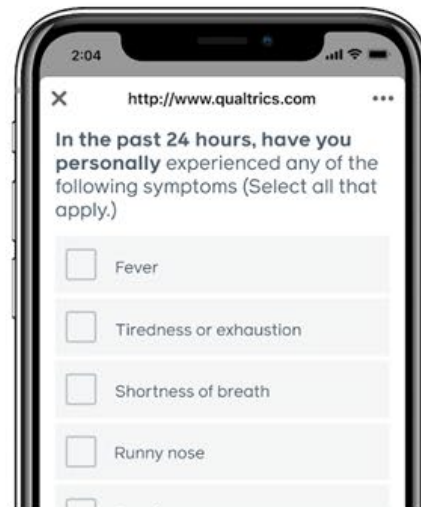
Global UMD CTIS Survey

“Do you personally know anyone in your local community who is sick with a fever and either a cough or difficulty breathing?”

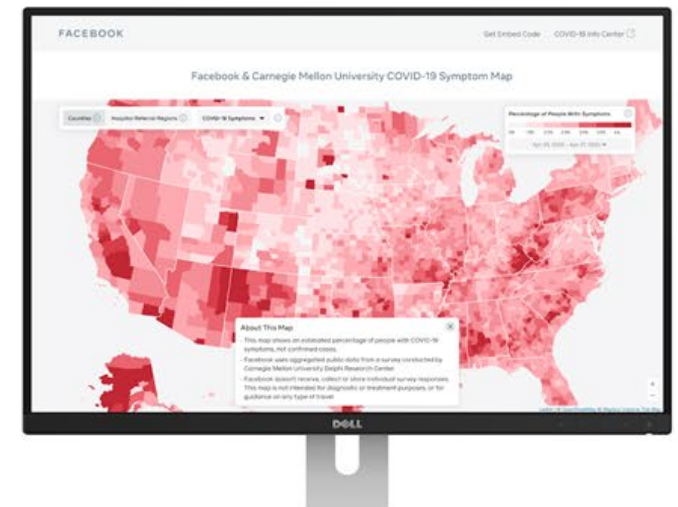
1 Who's Taking the Survey

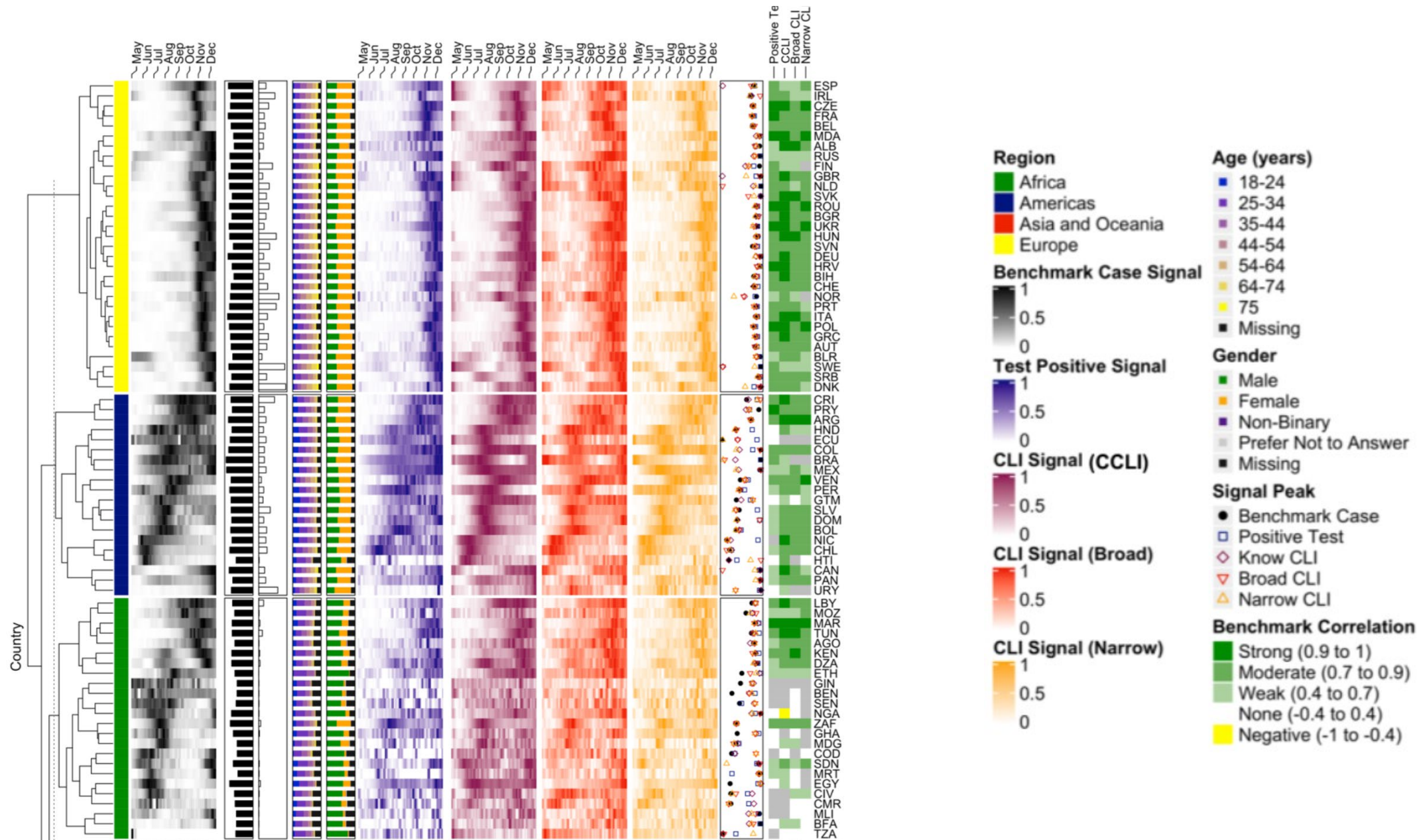


2 How the Survey Works



3 Using the Survey Data





Attitudes – Values - Opinions

Distributions - Intersectionality

The General Social Survey

The General Social Survey (GSS) is a nationally representative survey of adults in the United States conducted since 1972. The GSS collects data on contemporary American society in order to monitor and explain trends in opinions, attitudes and behaviors. The GSS has adapted questions from earlier surveys, thereby allowing researchers to conduct comparisons for up to 80 years.

The GSS contains a standard core of demographic, behavioral, and attitudinal questions, plus topics of special interest. Among the topics covered are civil liberties, crime and violence, intergroup tolerance, morality, national spending priorities, psychological well-being, social mobility, and stress and traumatic events.

Altogether, the GSS is the single best source for sociological and attitudinal trend data covering the United States. It allows researchers to examine the structure and functioning of society in general, as well as the role played by relevant subgroups and to compare the United States to other nations.

The GSS aims to make high-quality data easily accessible to scholars, students, policy-makers, and others, with minimal cost and waiting.

The GSS has carried out an extensive range of methodological research designed both to advance survey methods in general and to insure that the GSS data are of the highest possible quality. In pursuit of this goal, more than 130 papers have been published in the GSS Methodological Reports series.

International Social Survey Program

The ISSP, a cross-national collaboration conducting scientific surveys on diverse topics relevant to social science, evolved out of bilateral collaboration between NORC and the German organization Zentrum für Umfragen, Methoden, und Analysen (ZUMA; now part of GESIS-Leibniz Institute of the Social Sciences). Starting in 1982, each organization devoted a small segment of their national surveys, ALLBUS and GSS, to a common set of questions. The ISSP was formally established in 1984 by Australia, Germany, Great Britain, and the United States, and it now has 42 member countries across five continents and collects data in 70 countries. Each country's designated ISSP institution may decide what survey vehicle to field for the ISSP module each year, as long as data collection follows an approved methodology. As the only U.S. member of the ISSP, NORC actively participates in ISSP's international network, working to establish a framework for international cooperation that promotes measurement consistency

Need Help? ▶

Been Asked to Participate?

Has NORC contacted you to participate in the General Social Survey? If so, be sure to check out our Survey Participants page to learn more about the GSS, how your responses will be used and why your voice matters!

[Respond to the GSS](#)

GSS in the News

[Advisory: General Social Survey Has Created Social Media Archive, a New Source for Public Opinion Data](#)

NORC at the University of Chicago is giving social scientists and other researchers an easy way to fold social media conversations into their research projects with the launch of its General Social Media Archive. The Archive complements NORC's long-running, highly influential General Social Survey (GSS).

✳ [NORC.org](#) | December 14, 2022

[U.S. Why This Economic Boom Can't Lift](#)

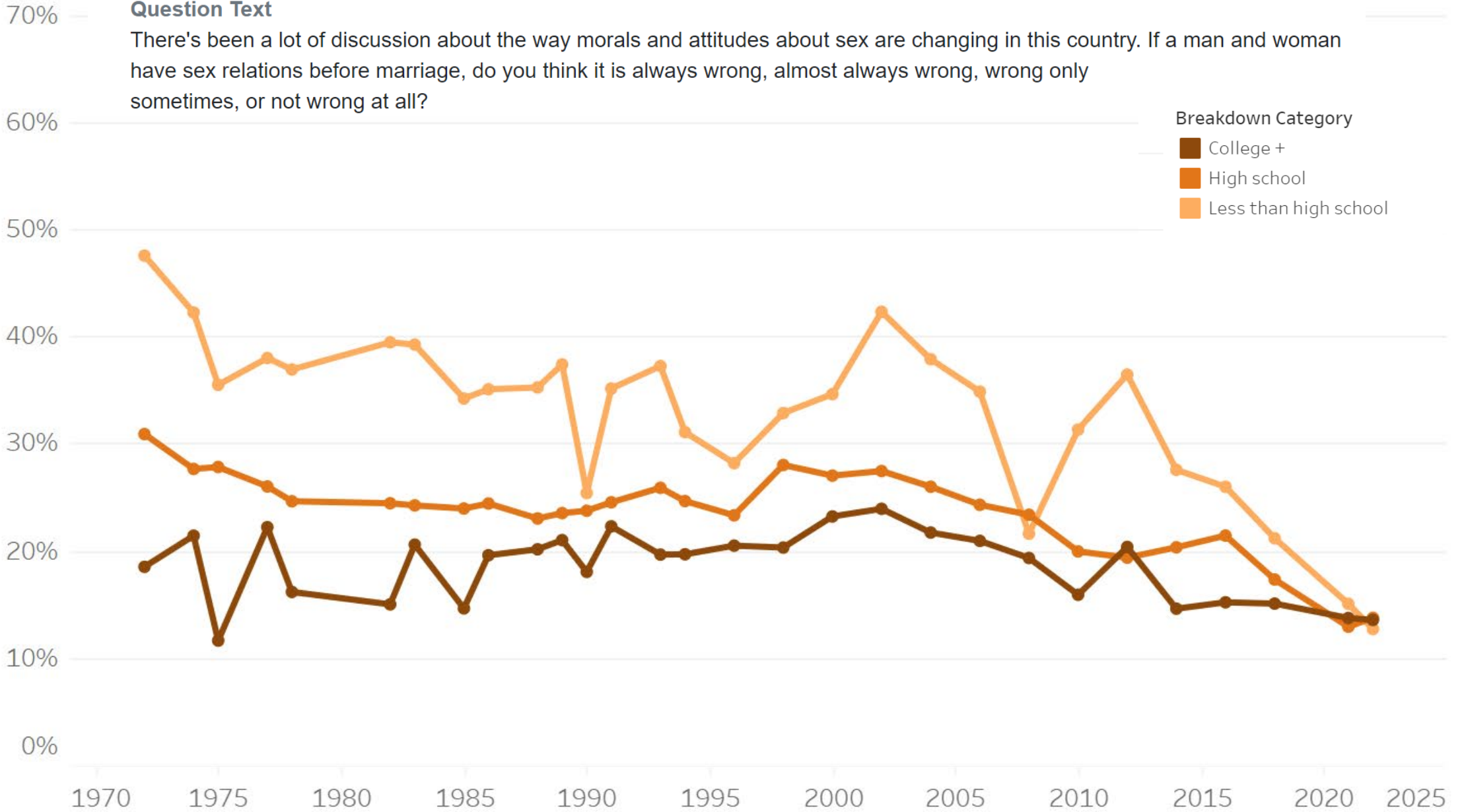


Question Text

There's been a lot of discussion about the way morals and attitudes about sex are changing in this country. If a man and woman have sex relations before marriage, do you think it is always wrong, almost always wrong, wrong only sometimes, or not wrong at all?

Percent of Population

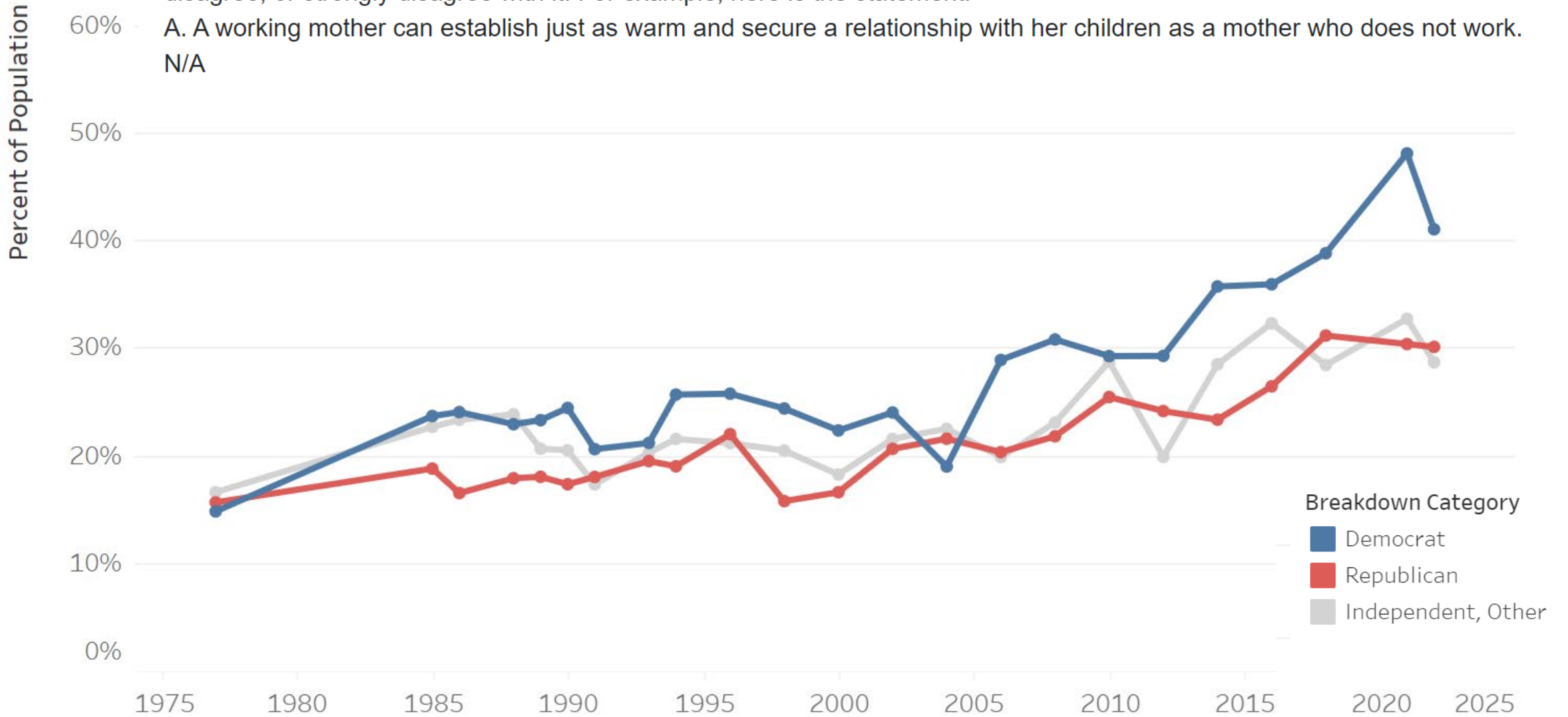
- Breakdown Category
- College +
 - High school
 - Less than high school



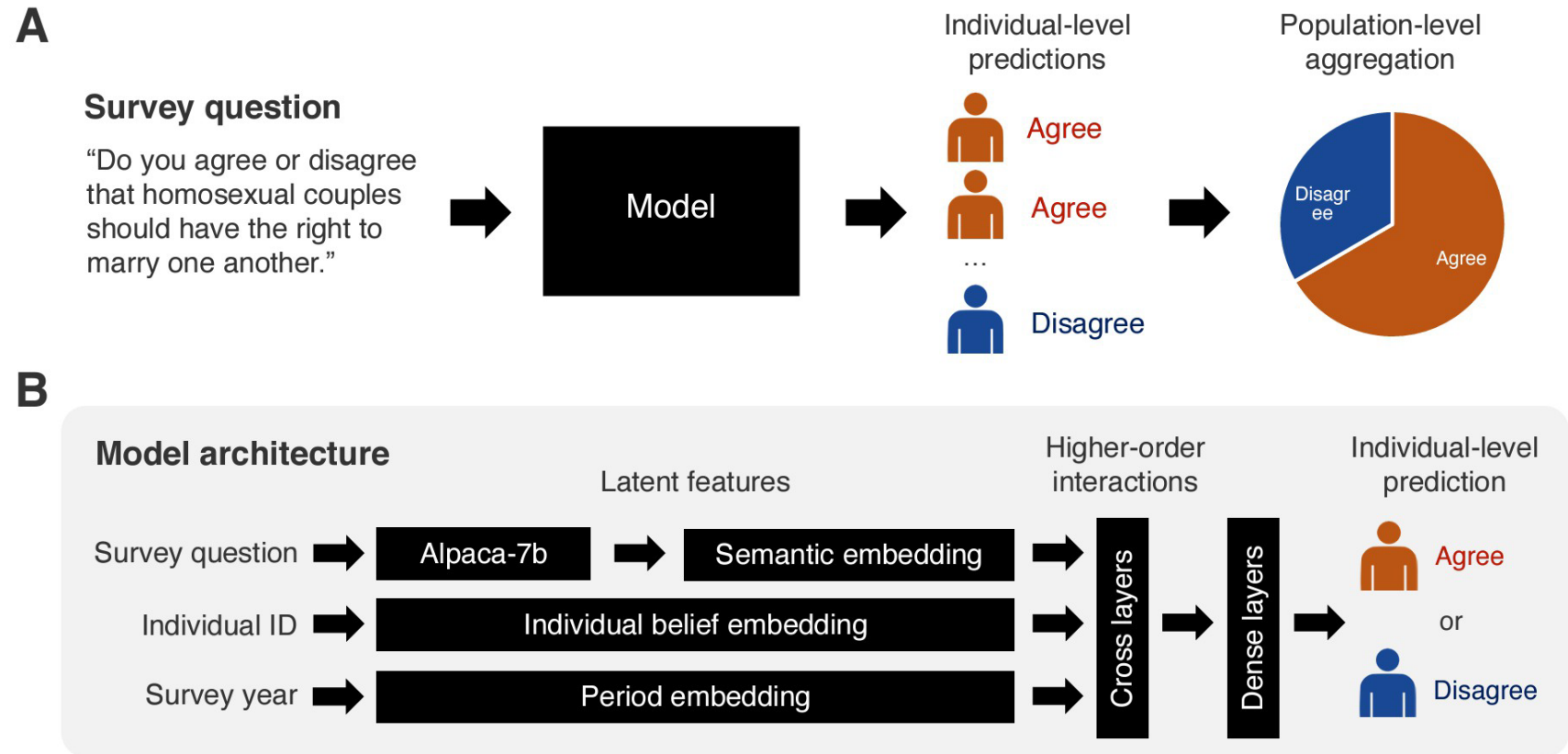
Question Text

Now I'm going to read several more statements. As I read each one, please tell me whether you strongly agree, agree, disagree, or strongly disagree with it. For example, here is the statement:

A. A working mother can establish just as warm and secure a relationship with her children as a mother who does not work.
N/A



Kim, J., Byungkyu, L., (2023, Nov 11). *AI-Augmented Surveys: Leveraging Large Language Models and Surveys for Opinion Prediction* <https://arxiv.org/abs/2305.09620>



DATA: 68,846 individuals' responses to 3,110 questions collected for 33 repeated cross-sectional data between 1972 and 2021 for fine-tuning the LLMs. Retrieved text content of GSS survey questions from GSS data explorer

Figure 2: An overview of our methodological framework. In Panel A, we use survey weights when aggregating individual-level prediction into population-level estimates to account for potential sampling bias. In Panel B, individual belief and period embeddings are initially randomly assigned but optimized during the fine-tuning process using dense and cross layers. Semantic embedding, initially estimated by pre-trained LLMs (e.g., Alpaca-7b), is also optimized during the fine-tuning stage.

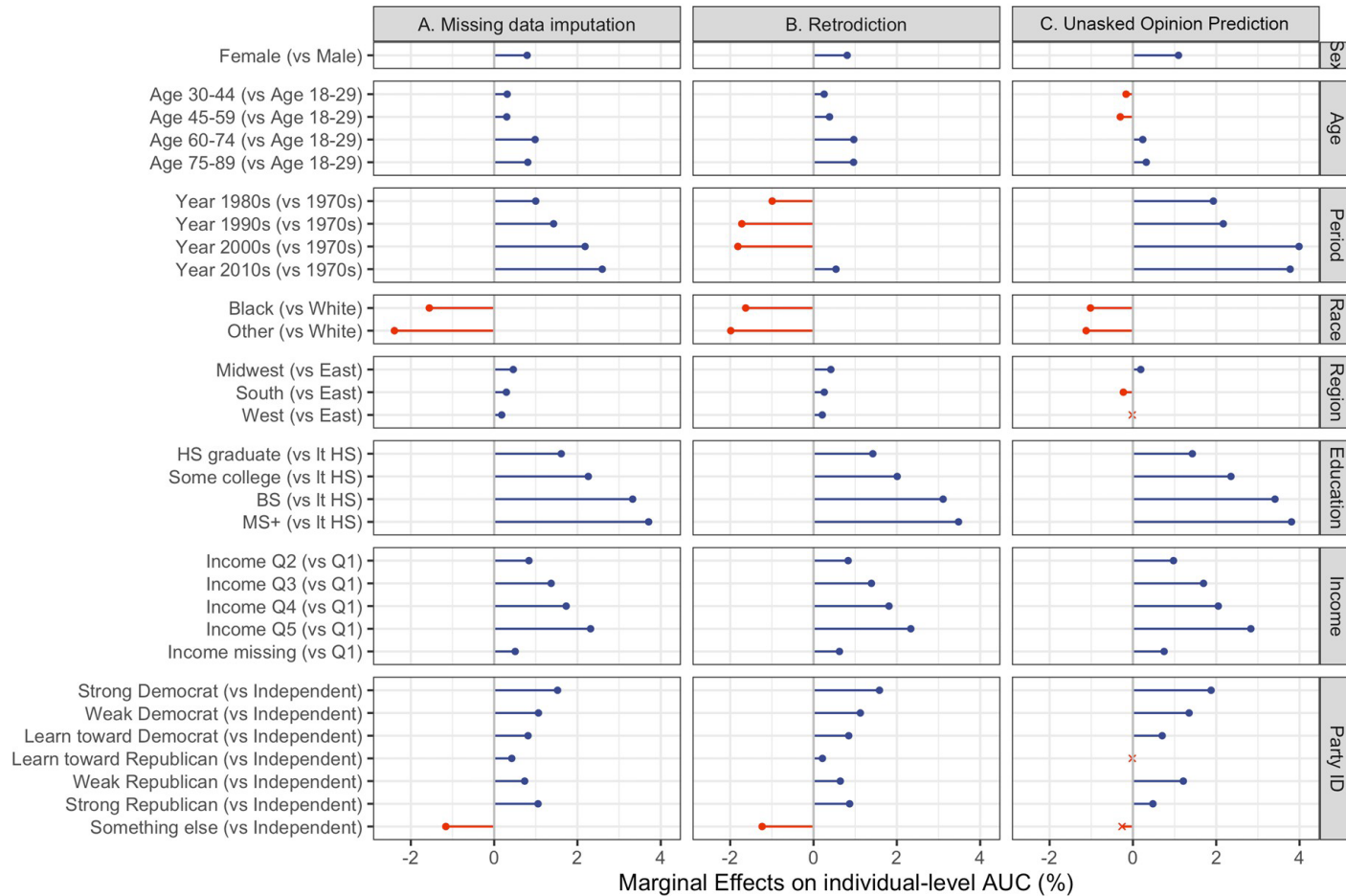


Figure 5: Coefficient plots from OLS regression models predicting individual-level AUC across three different types of missing response prediction. A higher AUC value indicates

For instance, rather than asking the same ten questions to a thousand participants, pollsters can disseminate twenty questions among the same thousand participants, each answering ten questions, and employ the model to **infer individual responses to the remaining ten unasked questions.** On the other hand, given our model’s remarkable ability to mimic human responses, even including biases, researchers can use it to **refine their survey questions by systematically examining characteristics of questions that cannot be accurately predicted (e.g., poor question wording).**

Kim, J., Byungkyu, L., (2023, Nov 11). *AI-Augmented Surveys: Leveraging Large Language Models and Surveys for Opinion Prediction* <https://arxiv.org/abs/2305.09620>

Measuring preferences

Is difficult but can be done.

We see the move reasonably with changes in environment.

Dimensions of Attitudes (Gallup 1947)

Step 1. *Filter Question*: Will you tell me what a 'filibuster in Congress' means to you?

Step 2. *Open Question*: What if anything, should Congress do about filibusters?

Step 3. *Dichotomous Question*: It has been suggested that the Senate change its rules so that a simple majority can call for an end to discussion instead of a two thirds majority as is now the case. **Do you approve or disapprove of this change?**

Step 4. *Reasons Why*: Why do you feel this way?

Step 5. *Intensity*: How strongly do you feel about this - very strongly, fairly strongly or not at all strongly?



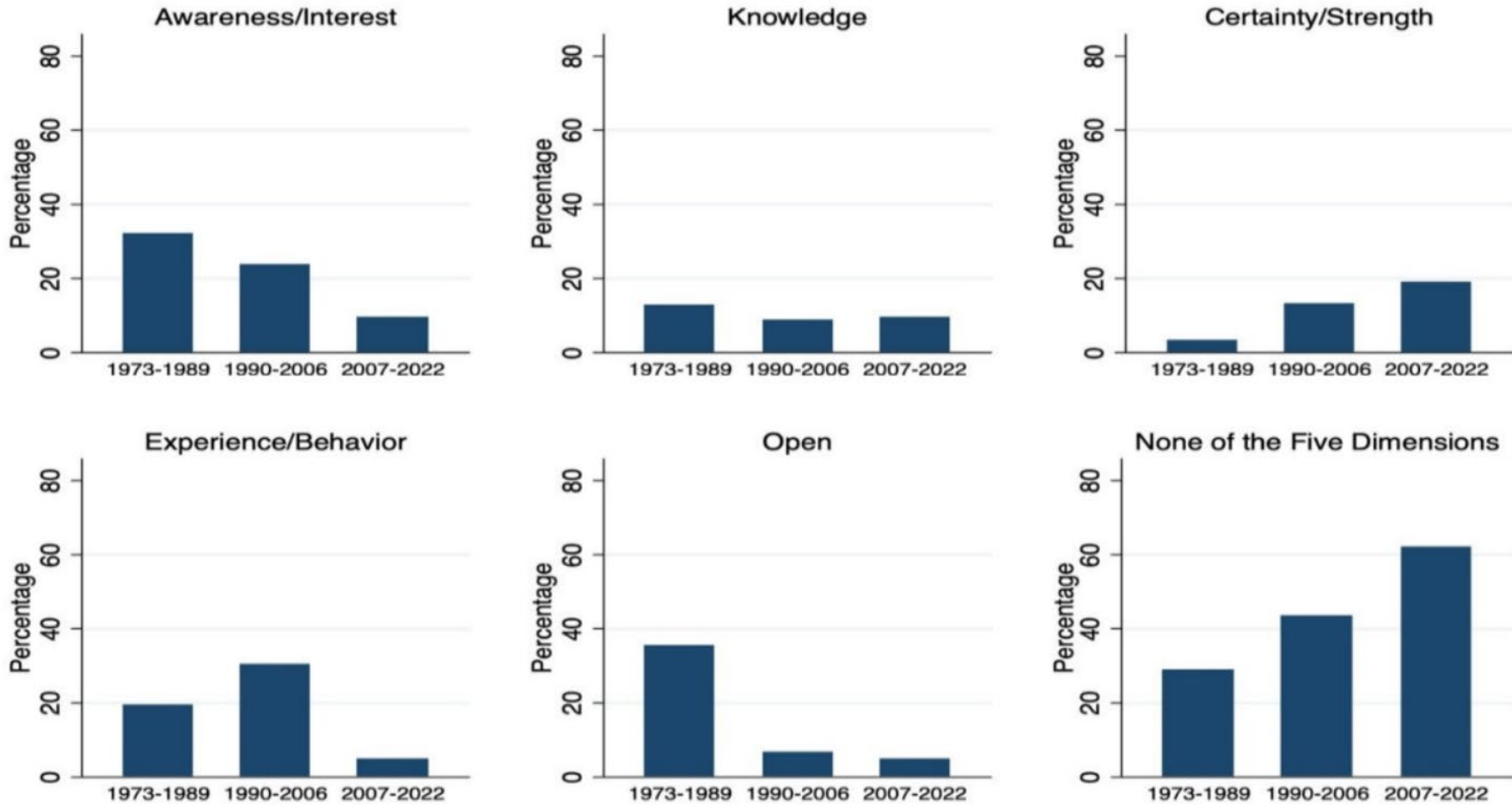
Stanley Presser
UMD



Sofia Jamie
UC Irvine



Dimensions Measured in Polls by Years



N: (1973-1989=31); (1990-2006=46); (2007-2022=21)

Context matters

Gerdon, Nissenbaum, Bach,
 Kreuter & Zins. 2021.
 Harvard Data Science
 Review
<https://doi.org/10.1162/99608f92.edf2fc97cc-by-4.0>

- Recipient:
- Public authority
 - Company

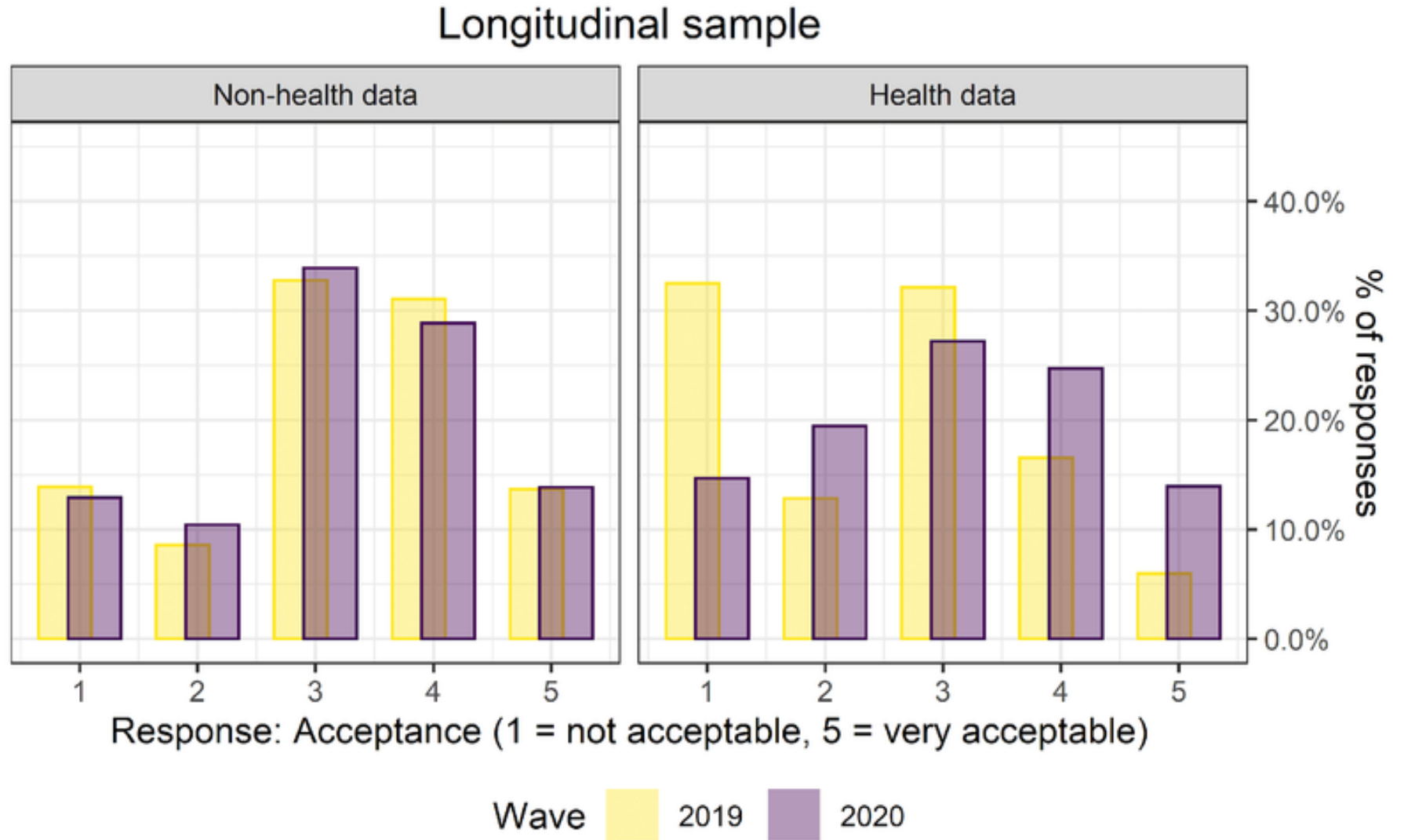
- Data type:
- Health: Sensors on a smartphone collect data on the health condition
 - Location: Smartphones collect location data during car rides
 - Energy: Intelligent power meters collect data on the energy consumption

Sensors installed on a smartphone collect data on the health condition of the holders (e.g., heart rate). With consent of the holder, these data are transmitted to a public authority. The public authority uses these data to detect the spread of infectious diseases in the population early and to develop solutions to their containment. The data are safe, anonymous, and protected from misuse.

Depending on data type:	Private purpose	Public purpose
Health	... personal recommendations on health behavior	...to detect outbreaks of infectious diseases early and to develop solutions to their containment.
Location	... personal recommendations on driving behavior and routes	...to develop improvements of the local infrastructure
Energy	... personal recommendations on the optimization and reduction of the own energy consumption	... to develop a more efficient energy distribution system

Context matters

Gerdon, Nissenbaum, Bach,
Kreuter & Zins. 2021.
Harvard Data Science Review
<https://doi.org/10.1162/99608f92.edf2fc97cc-by-4.0>



Distributions of Values

Data \neq Data

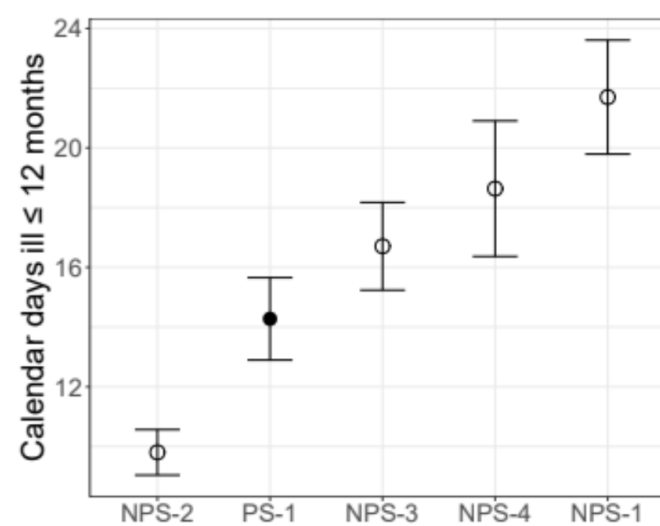
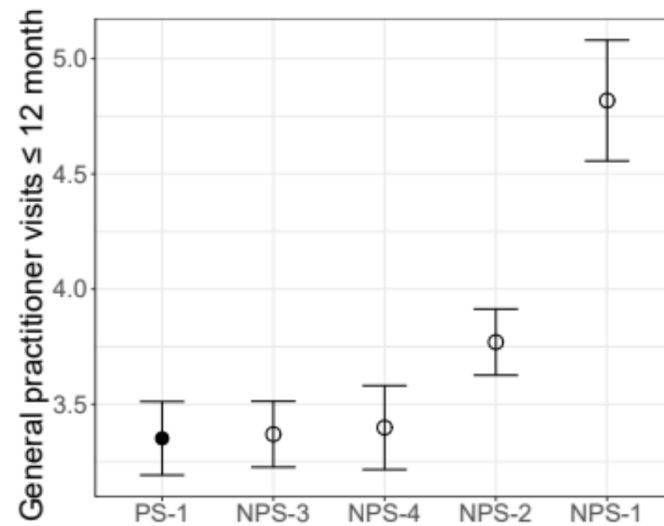
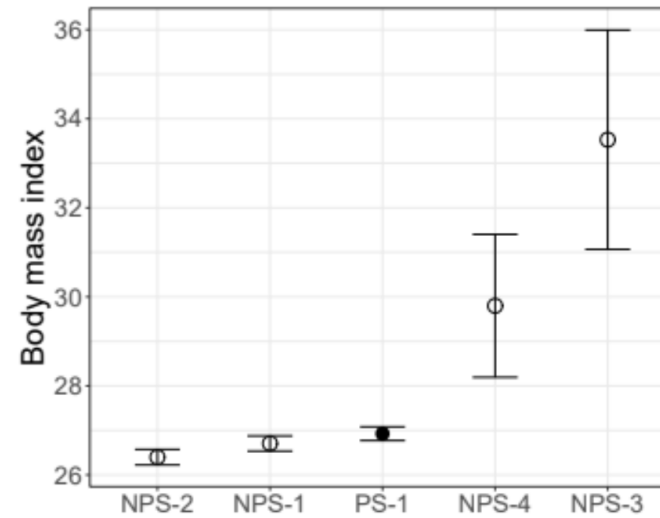
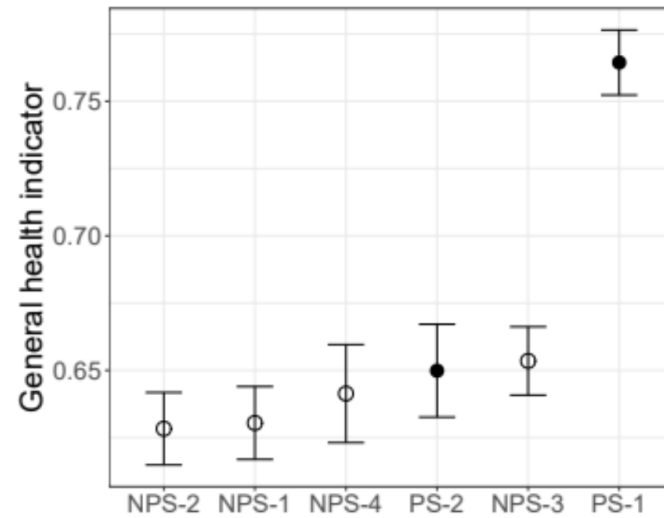
Health Estimate Differences Between Six Independent Web Surveys: Different Web Surveys, Different Results?

Rainer Schnell^{1*} and Jonas Klingwort²

¹Research Methodology Group, University of Duisburg-Essen, 47057 Duisburg, Germany

²Department of Research & Development, Statistics Netherlands (CBS), CBS-weg 11,
PO Box 4481, 6401 CZ Heerlen, the Netherlands

*To whom correspondence should be addressed; E-mail: rainer.schnell@uni-due.de



Abstract: Most general population web surveys are based on online panels maintained by commercial survey agencies. However, survey agencies differ in their panel selection and management strategies. Little is known if these different strategies cause differences in survey estimates. This paper presents the results of a survey

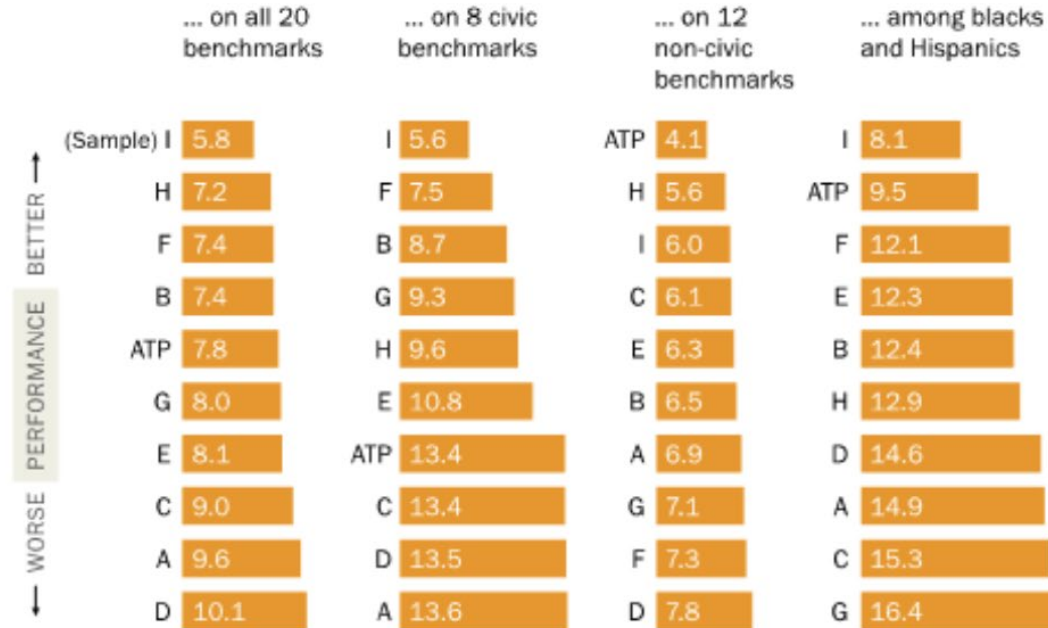
Excui

- Probal

Notable differences in data quality across online samples

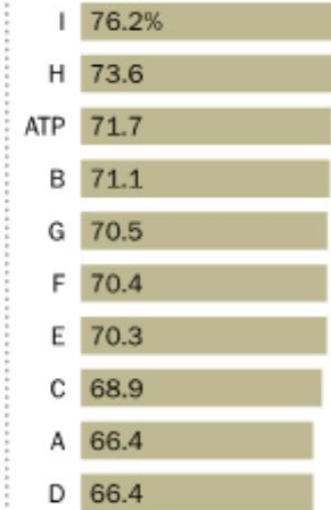
Average estimated bias in benchmarking analysis ...

Values for each sample represent the average of the absolute differences between the population benchmarks and weighted sample estimates



Average % correctly classified in regressions

How well regression models from online samples predict outcomes on benchmark samples (average across four outcomes)



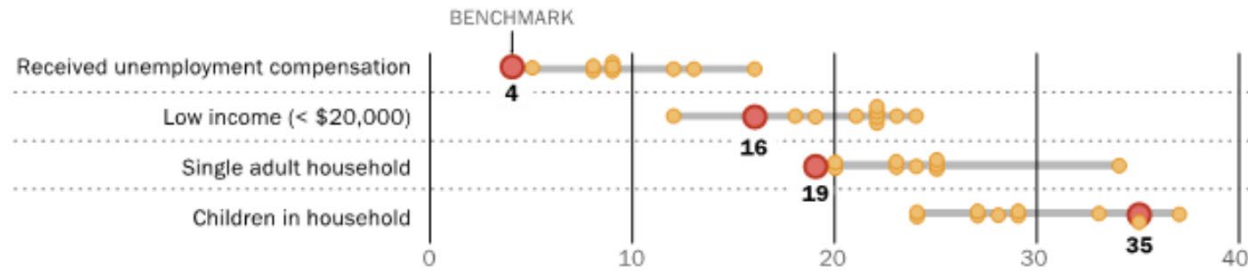
Note: Black and Hispanic averages exclude the driver's license benchmark which is not available for racial or ethnic subgroups. See Appendix D for details on individual benchmark items.

Source: Pew Research Center analysis of nine online nonprobability samples and the Center's American Trends Panel data. See Appendix A for details. "Evaluating Online Nonprobability Surveys."

PEW RESEARCH CENTER

Online samples tend to display a distinct socioeconomic profile

Weighted % of adults in online samples that belong to each category compared to federal benchmarks

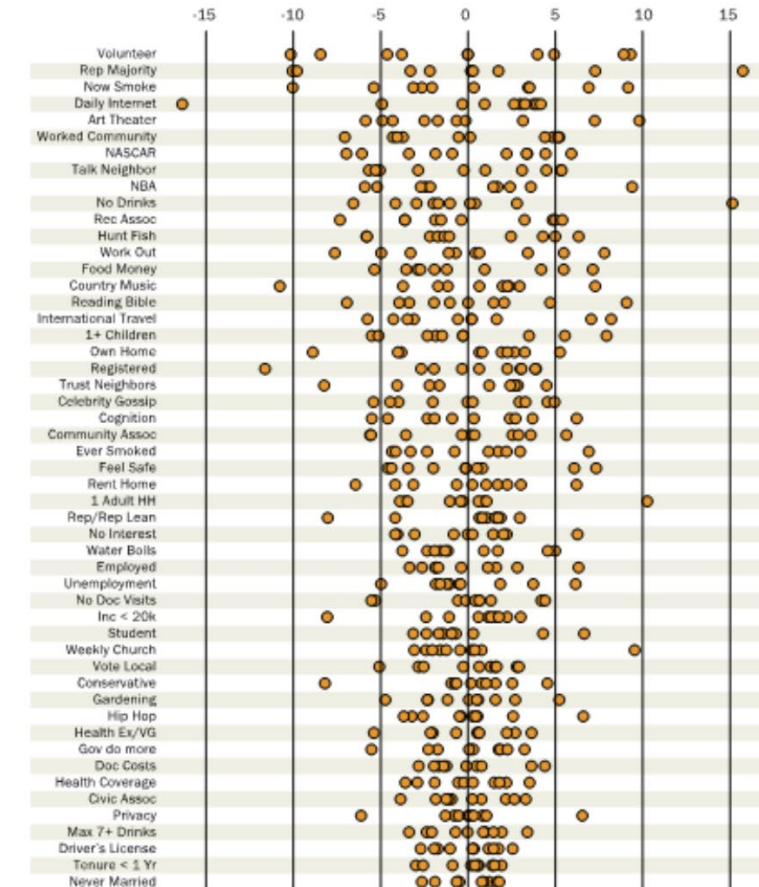


Source: 2015 Current Population Survey Annual Social and Economic Supplement; 2014 American Community Survey; Pew Research Center analysis of nine online nonprobability samples and the Center's American Trends Panel data. See Appendix A for details. "Evaluating Online Nonprobability Surveys."

PEW RESEARCH CENTER

Estimates from different online samples run the gamut from highly variable to highly consistent

Deviation between each weighted estimate and the grand mean of estimates from all 10 online samples



Source: Pew Research Center analysis of nine online nonprobability samples and the Center's American Trends Panel data. See Appendix A for details. "Evaluating Online Nonprobability Surveys."

PEW RESEARCH CENTER



HUMAN FEEDBACK FROM
WHOM AND HOW



SUGGESTION FOR CHANGING
THE (PRE)-TRAINING DATA



DISCUSSION: HOW CAN WE
COLLABORATE?

Questions

- What level of aggregation is still acceptable?
- How much noise would be tolerate on the microdata?
- Which format would be useful?

Question Text

There's been a lot of discussion about the way morals and attitudes about sex are changing in this country. If a man and woman have sex relations before marriage, do you think it is always wrong, almost always wrong, wrong only sometimes, or not wrong at all?

- What value does value alignment have? Funding for good benchmark surveys?

AAPOR 80th Annual Conference

Reshaping Democracy's Oracle: Transforming Polls, Surveys, and the Measurement of Public Opinion in the Age of AI

May 14 - 16, 2025
St. Louis

AAPOR 80th Annual Conference

The AAPOR Annual Conference is the premier forum for the

[Call for Abstracts](#)

[Schedule at a Glance](#)

