AIR®
Advancing Evidence.
Improving Lives.

Population Estimates Based on Social Media Scraping

AAPOR Conference

Cong Ye, Principal Researcher
cye@air.org

AAPOR Conference| May 2022

Agenda

1. Introduction
2. Methodology
3. Results
4. Discussions

AIR®
Advancing Evidence.
Improving Lives.

Introduction

Opportunity and Challenges

Introduction

- Social media has become a major communication channel
  - Opportunity for studying district policies
- Compared with surveys
  - Faster
  - Little burden on districts
  - Less effort from researchers
- Challenges
  - Bias
  - Accuracy

AIR®
Advancing Evidence.
Improving Lives.

Methodology

Sampling, Scraping, Weighting, Topic Generating, Tagging, Estimating, and Benchmarking

Sampling

- A representative sample
  - Stratified by state (CA, TX, and other) and urbanicity
  - Systematic sample from a list sorted by district size, percent minority student, grade level served
  - 2,000 districts with 500 in CA and 500 in TX

Scraping

- Scrape for social media accounts
  - Start with district website information in CCD
  - If not available, search for the website
  - Identify the social media accounts on website and save to a NoSQL database
- Request tweets and Facebook posts
  - Data saved to a NoSQL database

Weighting

- Base weight
  - Inverse of selection probabilities
- Missing data adjustments
  - No account
  - Use pattern (number of twits/posts)

Topic Generating

- Data preparation
  - Google Translate's API to translate non-English contents
  - Replace special characters, lower case all words, as well as strip punctuation and symbols
  - Drop stop words – unimportant words such as "the", "is" and "and"
  - Tabulate word frequencies and review for additional cleaning
    - » Multiple iterations
- Lemmatization
  - Turn words to their root forms
    - » went to go, caring to care, children to child
  - Use word collocation model to identify two-word phrases (bigrams)

Topic Generating, Continued

- Latent Dirichlet Allocation (LDA)
  - Using the Mallet's implementation which takes a Gibbs sampling approach to modeling
  - Review results for coherence and decide the final model

Tagging

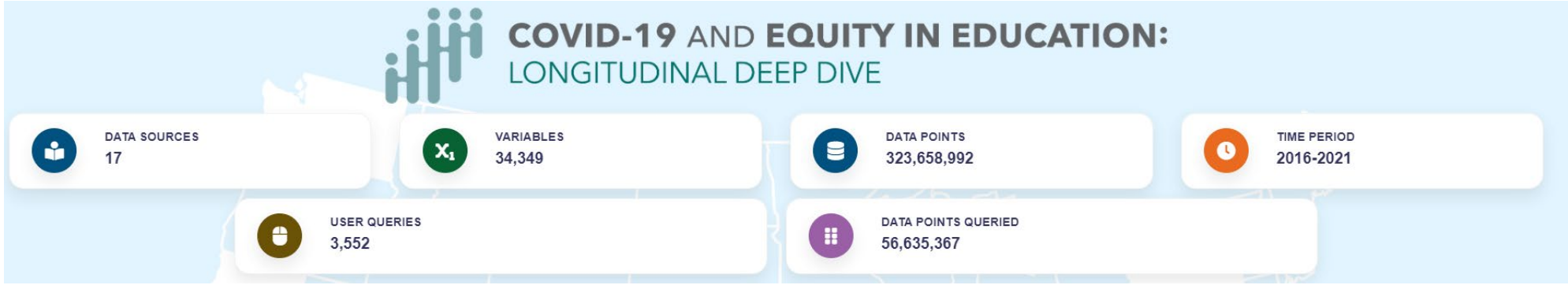
- For a given topic, generate statements
  - E.g., Mask is required, We require wearing a mask, Face covering is required
- Tag the district tweets/posts for the target statements
  - Probability tagging on lemmatized statement and tweet/post
  - Threshold setting based on review

Estimating

- Focus on the spring semester (January 2021 to May 2021)
  - Based on available benchmark data
- Account for differential social media use patterns
  - Weighted analysis
- Variance estimation
  - Account for stratification and finite population correction

Benchmarking

- U.S. School Closure & Distance Learning Database (osf.io/tpwqf)
  - Based on mobile data
- COVID-19 School Data Hub (covidschooldatahub.com)
  - Administrative data
- Data for some states from the two data sources are available in the CEE database (cee.airprojects.org)

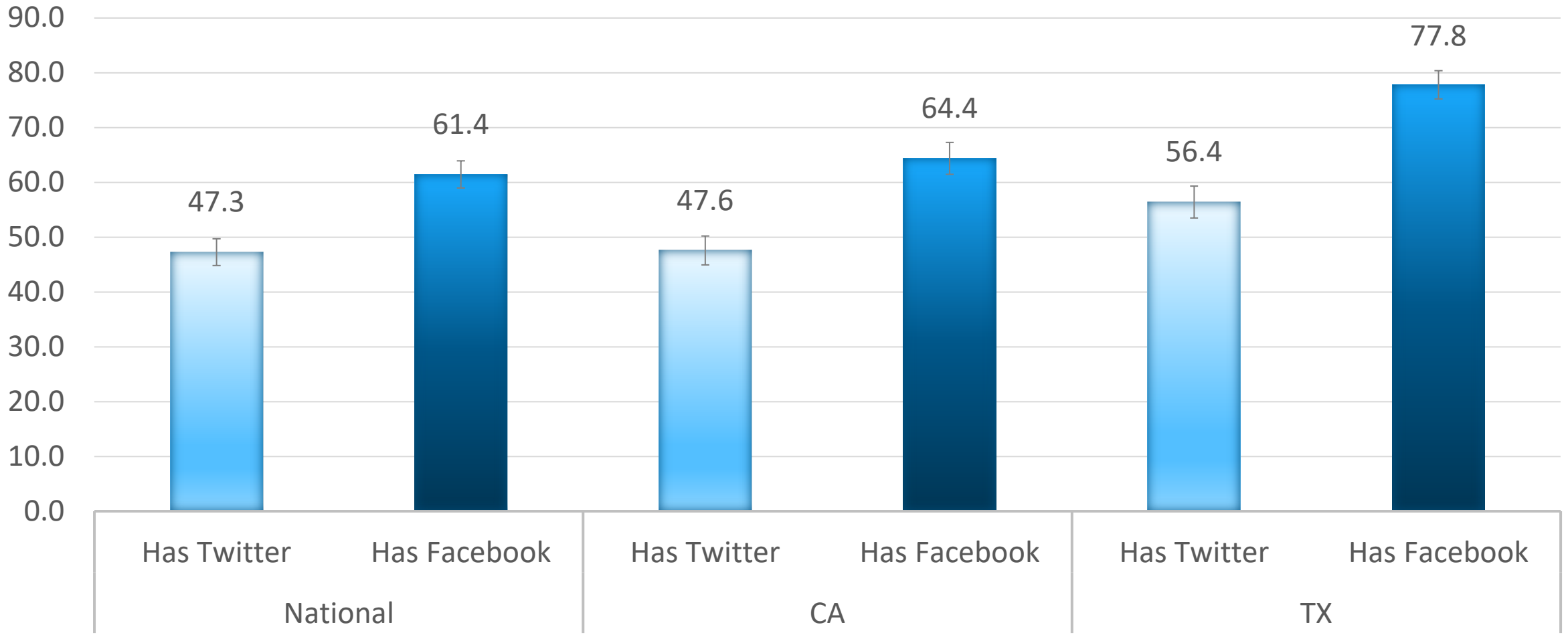


AIR®
Advancing Evidence.
Improving Lives.

Results

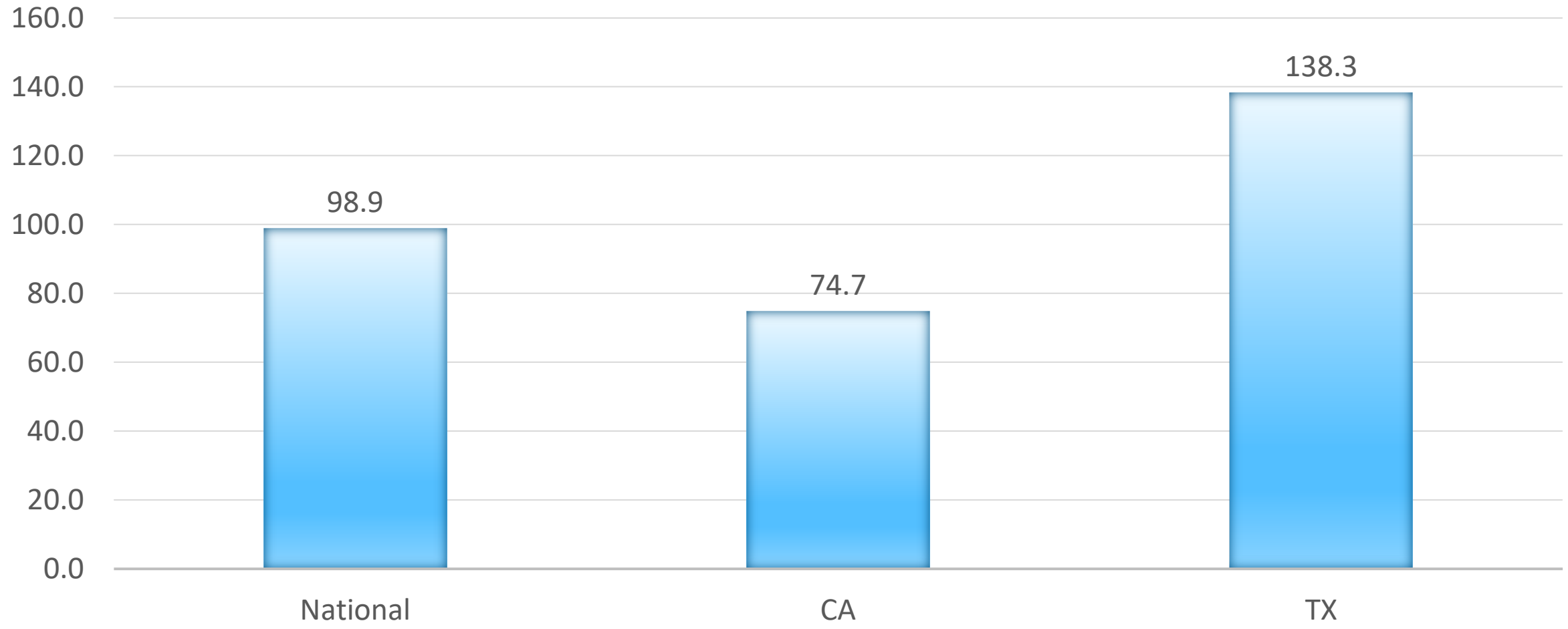
Accounts, Active Use, Topic Mentioned Compared with Benchmarks

Results – % Districts with Social Accounts



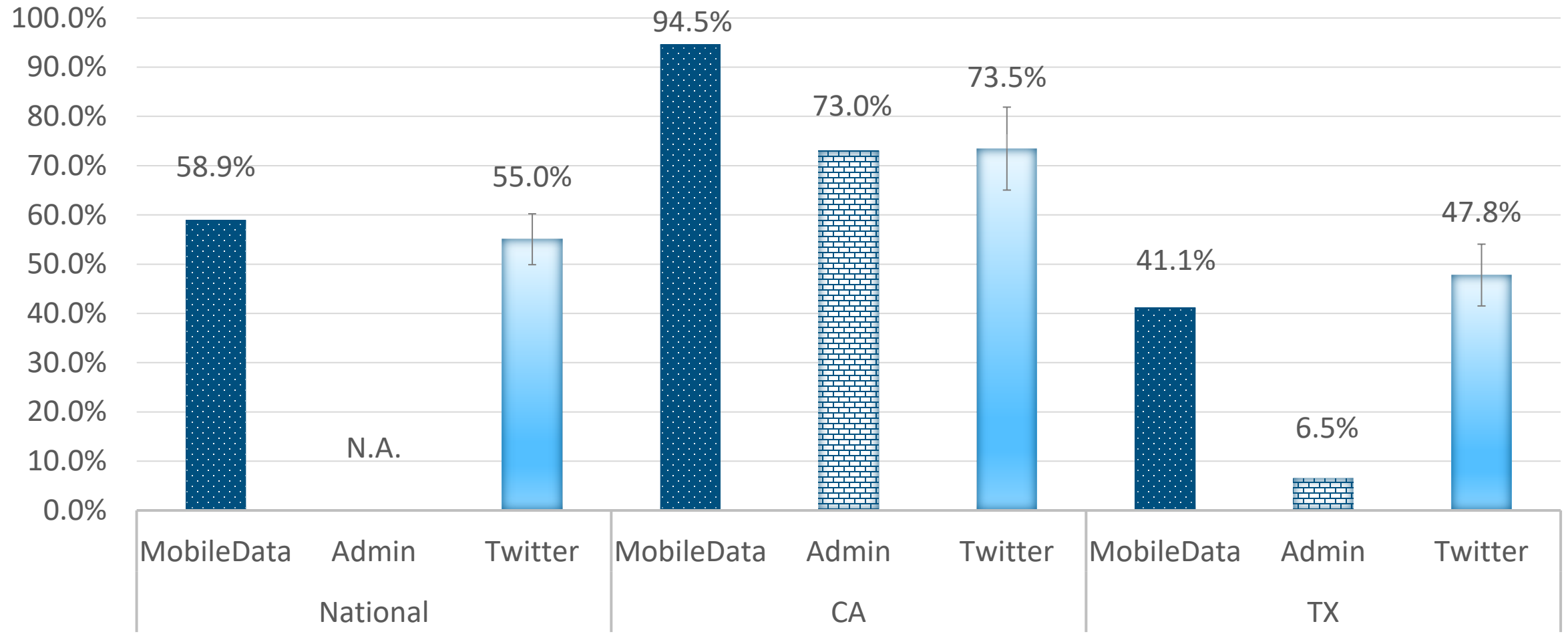
Note: Estimates based on results from web scraping in March 2022.

Results – # of Tweets from January to May 2021, Unweighted



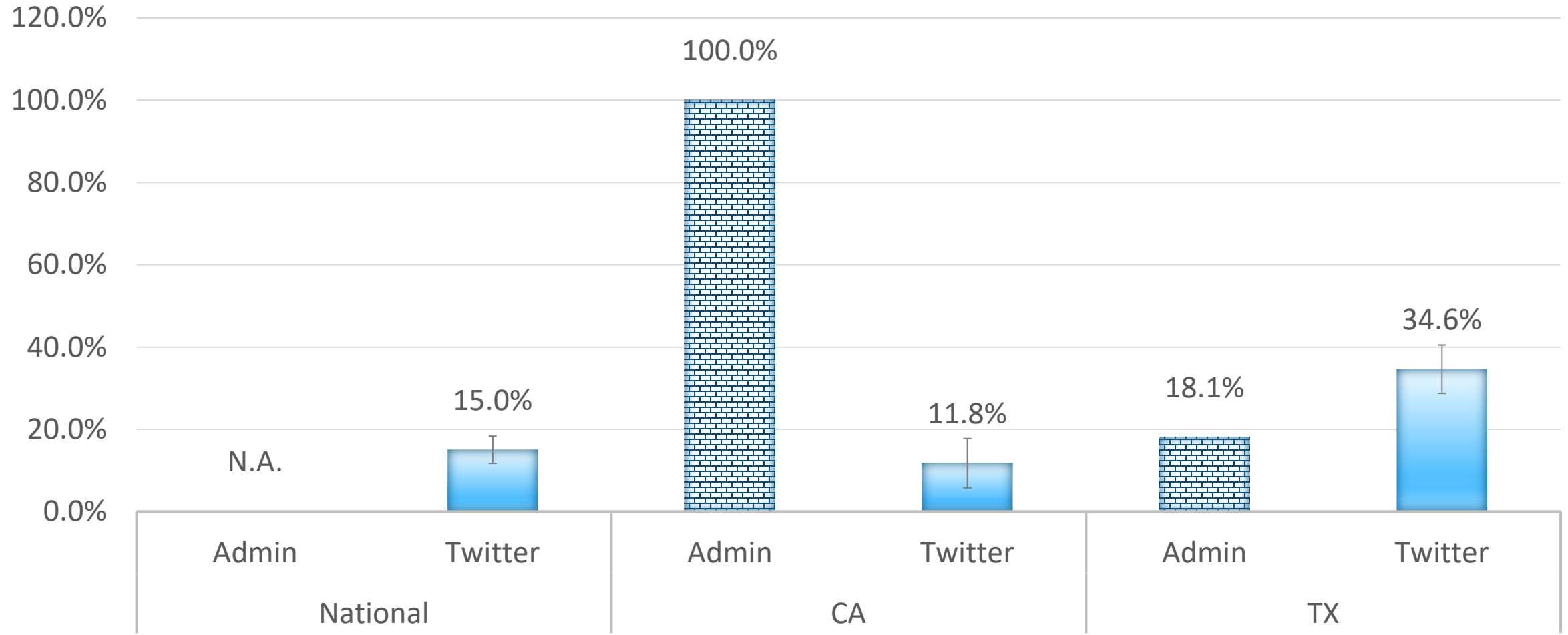
Note: Estimates based on results from web scraping in March 2022.

Results – Distance Learning



Note: N.A. – Not Available; Estimates based on results from web scraping in March 2022.

Results – Mask Requirements



Note: N.A. – Not Available; Estimates based on results from web scraping in March 2022.

AIR®
Advancing Evidence.
Improving Lives.

Discussions

Topic/Context Dependent, Coding of Pictures, Uncertainty in Probability-Based Tagging

Discussions

- Topic/context dependent
  - Some topics more likely to appear on social media than others
  - Important to understand the context
- Pictures/Videos
  - Coding
  - Alternative texts
- Probability-Based Tagging
  - Incorporate the uncertainty in variance estimation

AIR®
Advancing Evidence.
Improving Lives.

Cong Ye

Principal Researcher
+1.202.403.7168
cye@air.org