VERSTA RESEARCH

Finding Fraud in Public Polls

Employing Semantic Network-Based Methods for Identifying Fraud in Online Sampling

AAPOR 76th Annual Conference, May 11-14, 2021

Removing Fraudulent Respondents

The Speedster

- Speeding
- Inattentive
- Inconsistent
- Unengaged

The Fraudster

- Yes to everything
- Always be eligible
- Benefit optimization

The Bot

- Automated scripts
- Dark web tech
- Click farms

Problematic Respondents:

- ***Provide noise:*** inattentive, unengaged, respond randomly, or don't understand the survey
- ***Systematic bias:*** yea-saying, acquiescence bias, mischievous responding

Assessing Data Quality

Have Mischievous Responders misidentified sexual minority youth disparities in the National Longitudinal Study of Adolescent and Adult Health?

Archives of Sexual Behavior
May 5, 2017

Jessica Fish and Stephen Russel

Assessing the Risks to Online Polls From Bogus Respondents:

Pew Research Center
February 18, 2020

Courtney Kennedy, Nick Hatley, Arnold Lau, Andrew Mercer, Scott Keeter, Joshua Ferno, and Dorene Asare-Marfo

Teenagers "use" of Non-Existent Drugs: A study of false positives

Nordic Studies on Alcohol and Drugs
February 1, 2006

Hilde, Pape. and Elisabet E. Storvoll,

Large Impact of Small Differences

Gargling with bleach? Americans misusing disinfectants to prevent coronavirus, survey finds

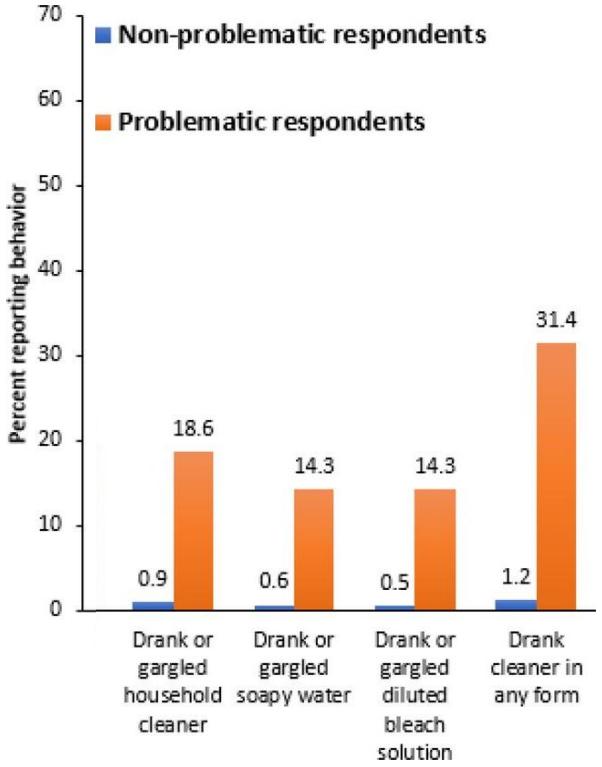
By Reuters Staff 2 MIN READ

FILE PHOTO: Bottles of Clorox bleach are displayed for sale on the shelves of a Wal-Mart store in Rogers, Arkansas June 4, 2009. REUTERS/Jessica Rinaldi

(Reuters) - More than a third of Americans misused cleaners and disinfectants to try to prevent infection by the coronavirus, according to a survey taken shortly after President Donald Trump publicly asked whether injecting such products could treat COVID-19.

Washing food with bleach, using household cleaning or disinfectant products on bare skin, and intentionally inhaling or ingesting these products were some of the most commonly reported "high-risk" practices in a May 4 online survey of 502 U.S. adults, the Centers for Disease Control and Prevention (CDC) reported.

- Over 90% of reports of ingesting bleach were made by problematic respondents



Did people really drink bleach to prevent COVID-19? A tale of problematic respondents and a guide for measuring rare events in survey data

Leib Litman, Zohn Rosen, Cheskie Rosenzweig, Sarah L. Weinberger-Litman, Aaron J. Moss, Jonathan Robinson. medRxiv 2020.12.11.20246694; doi: https://doi.org/10.1101/2020.12.11.20246694

Testing Presurvey Screening

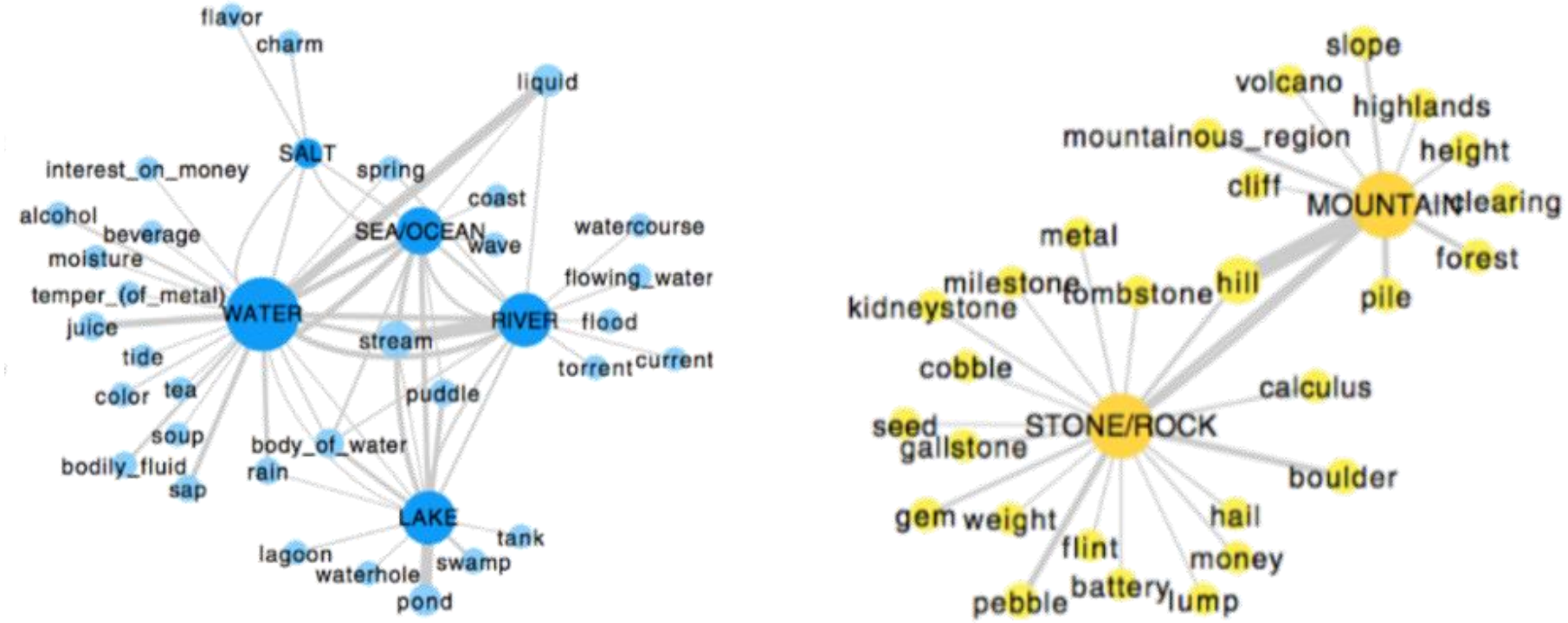
Evaluate performance of presurvey screening in classifying problematic respondents

- Three tests implemented alongside online surveys conducted in July, September, October 2020
- High quality, reputable sources who regularly supply online sample from opt-in panels for public opinion studies
- Evaluating performance in different sample/questionnaire/topical environments
- Respondents classified using presurvey screens, in-survey measures, and traditional postsurvey review

Semantic Network Model Approach

Goals:

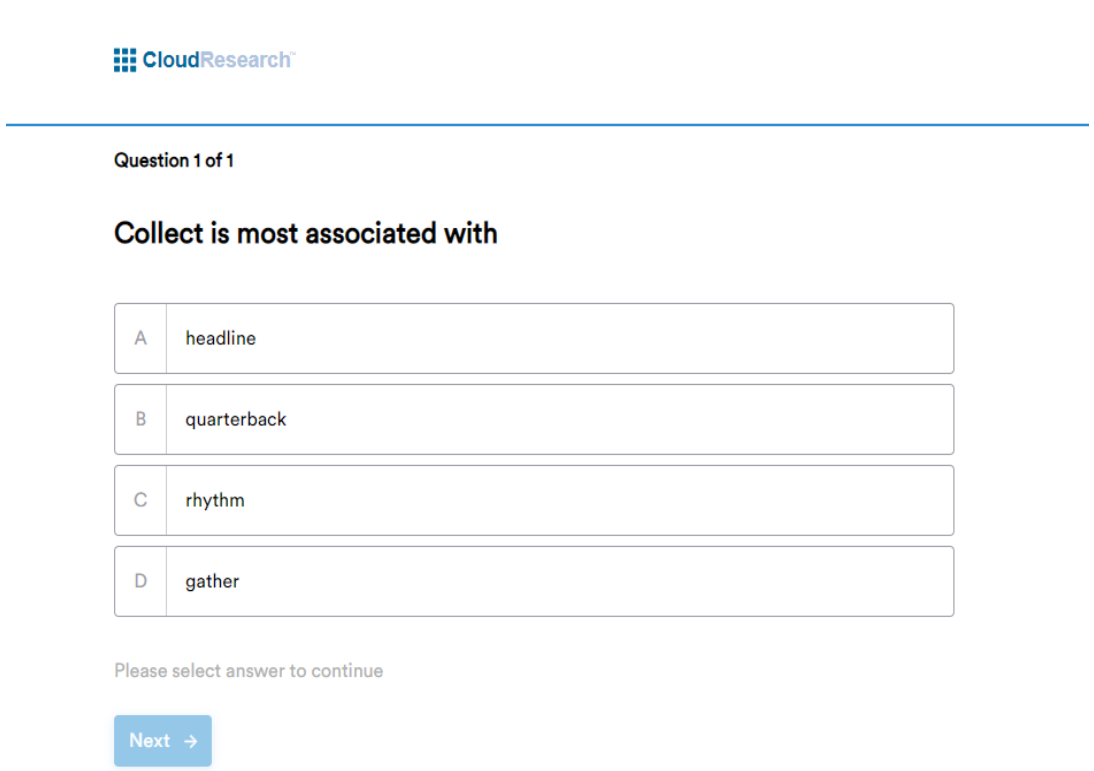
- *Scalability:* thousands of instrument items are needed to prevent fraud
- Difficulty levels based on objective, quantifiable, and adjustable criteria
  - Difficulty quantified based on word frequency and threshold adjustment

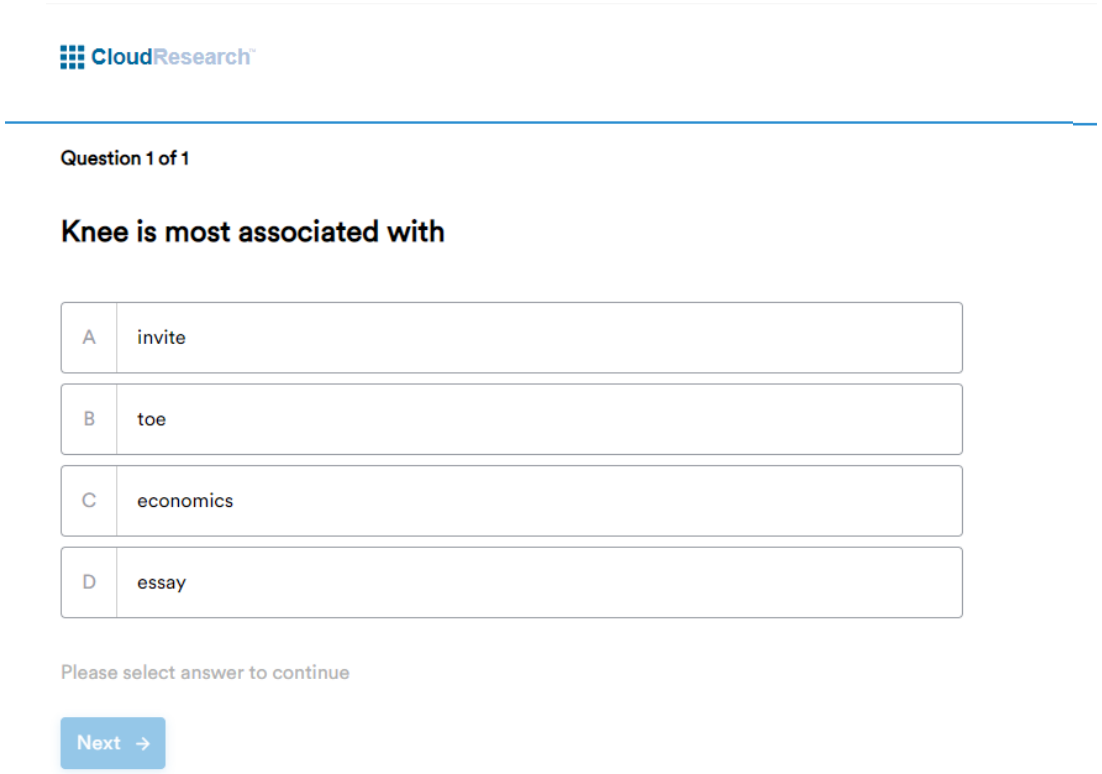


Youn, H., et al (2016). On the universal structure of human lexical semantics. *Proceedings of the National Academy of Sciences*, *113*(7), 1766-1771.

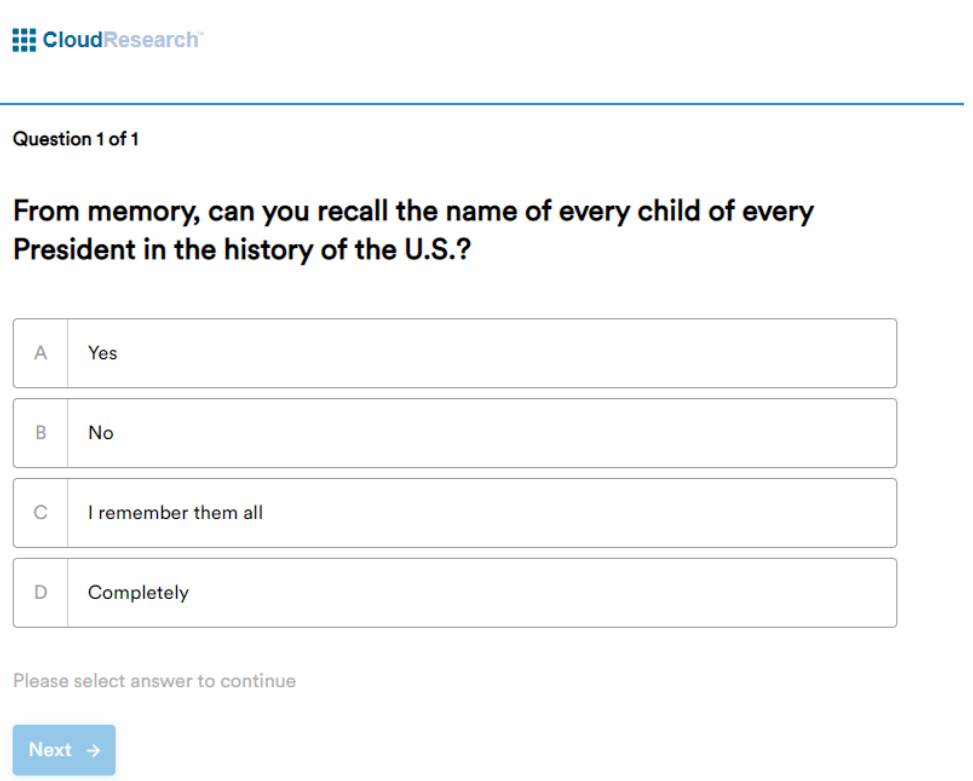
Semantic Network-Based Stimuli

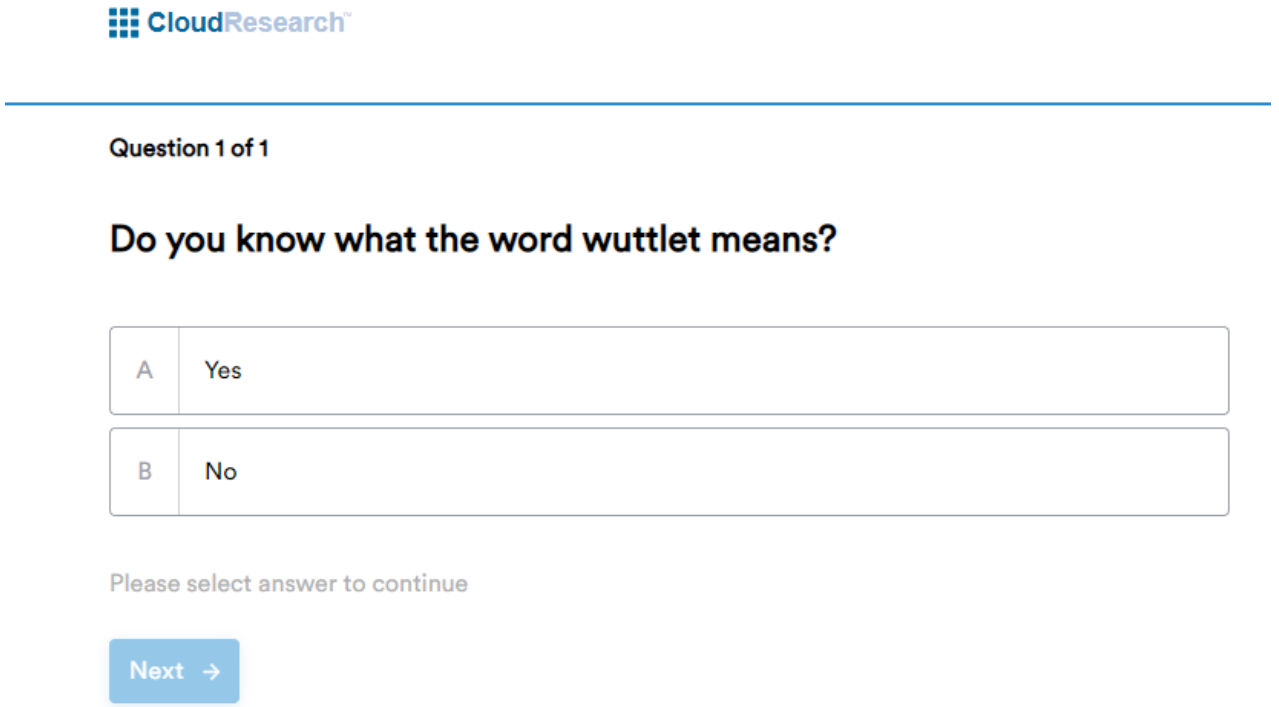
- Question creation and scoring based on associative network models
- Extensive testing, stabilization of optimal difficulty level



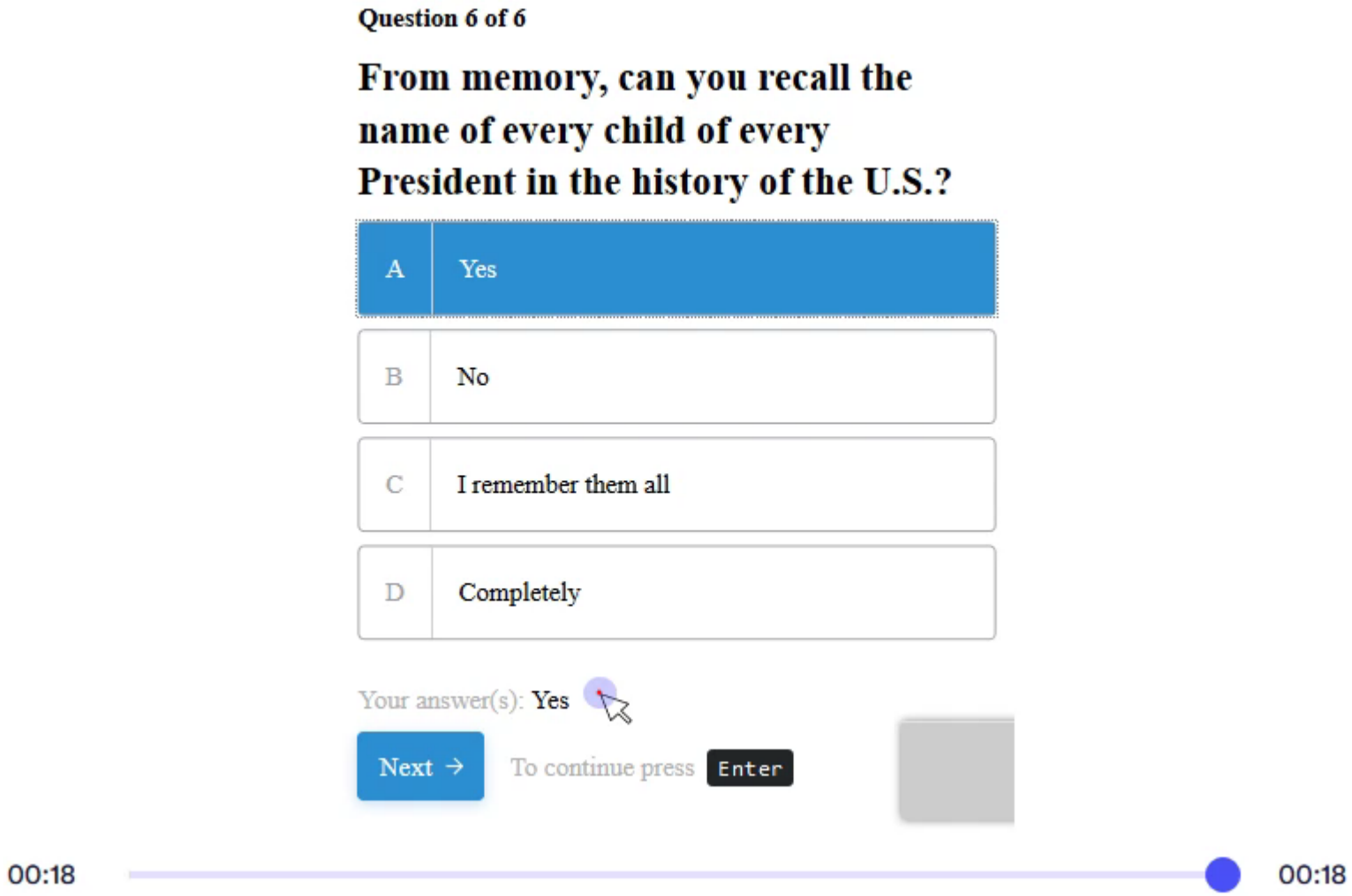


Yea-Saying Instrument Items

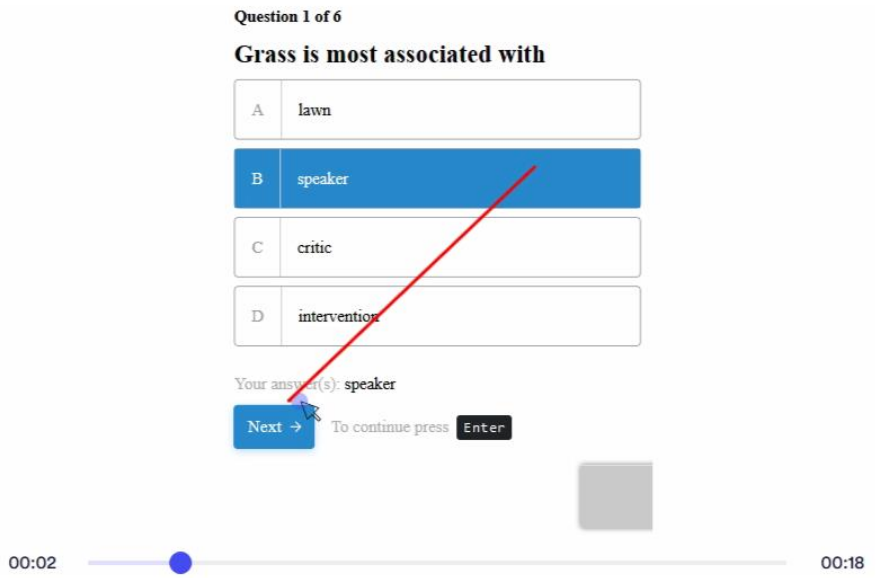


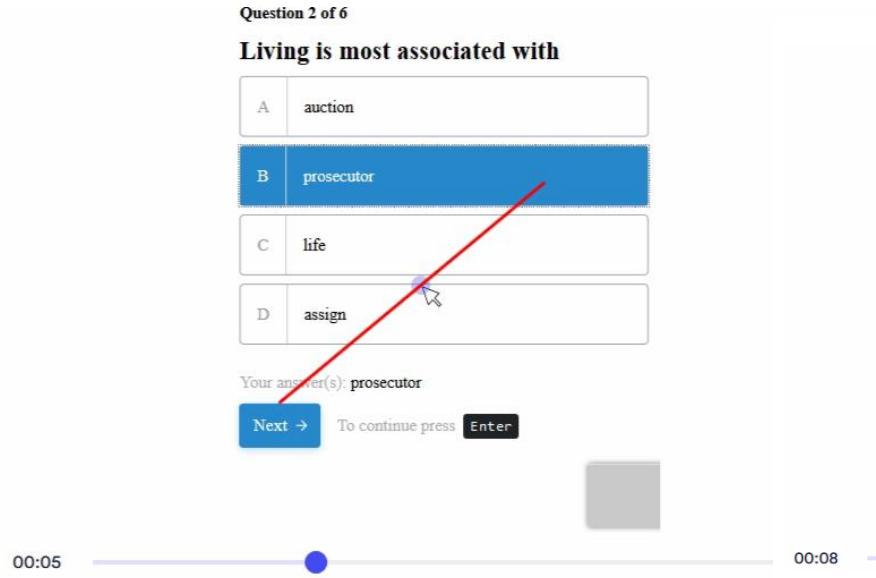


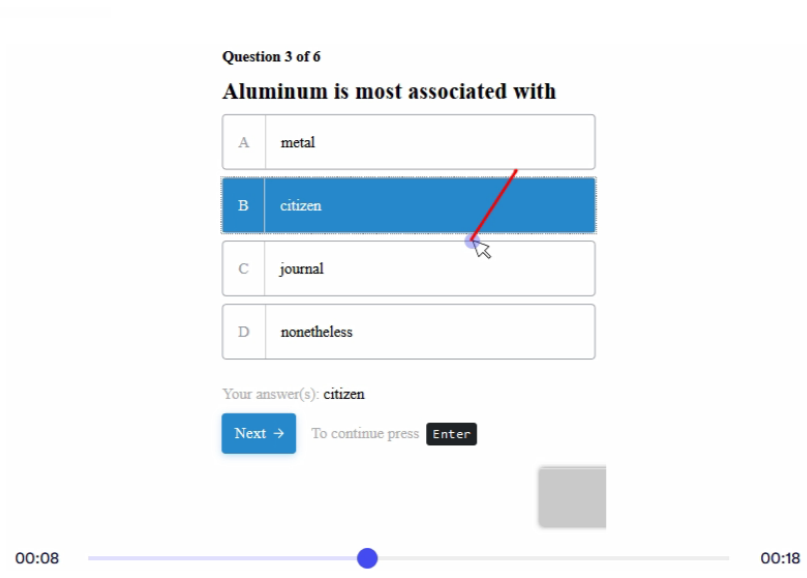
Actual Respondent Taking Presurvey Screener

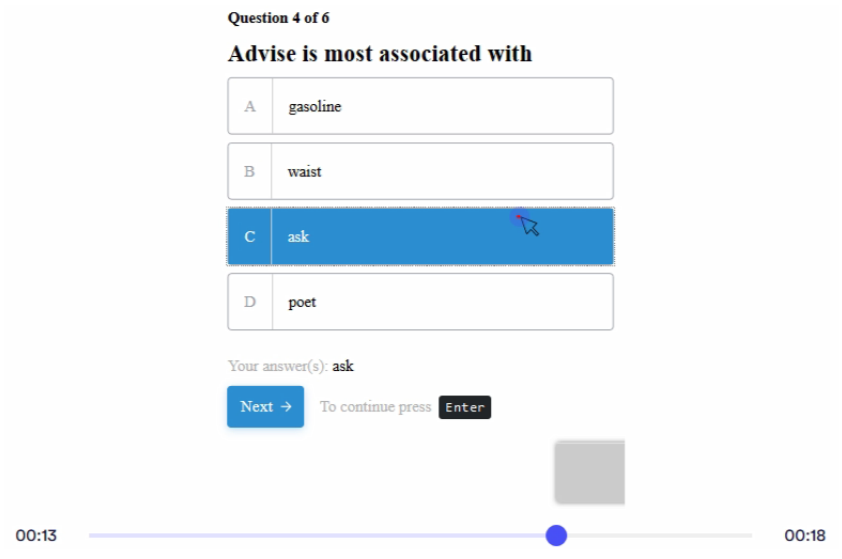


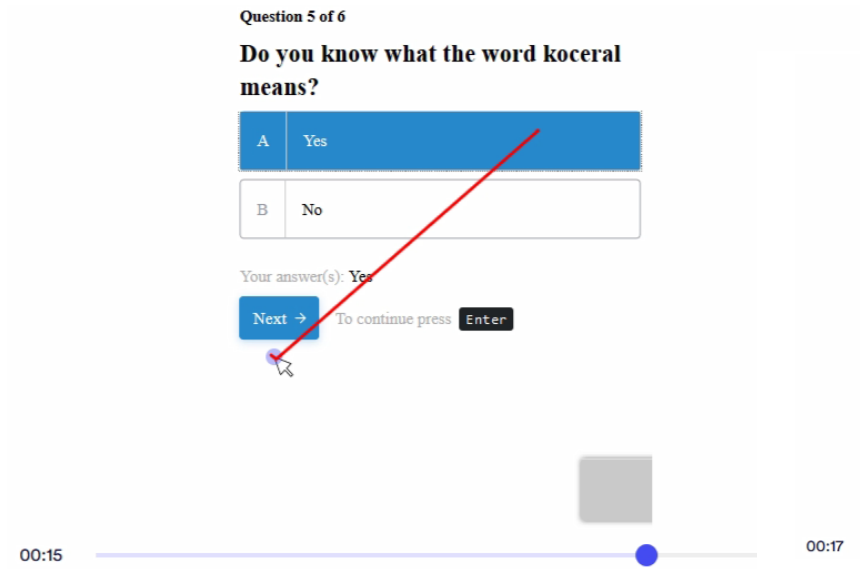
Actual Respondent Taking Presurvey Screener

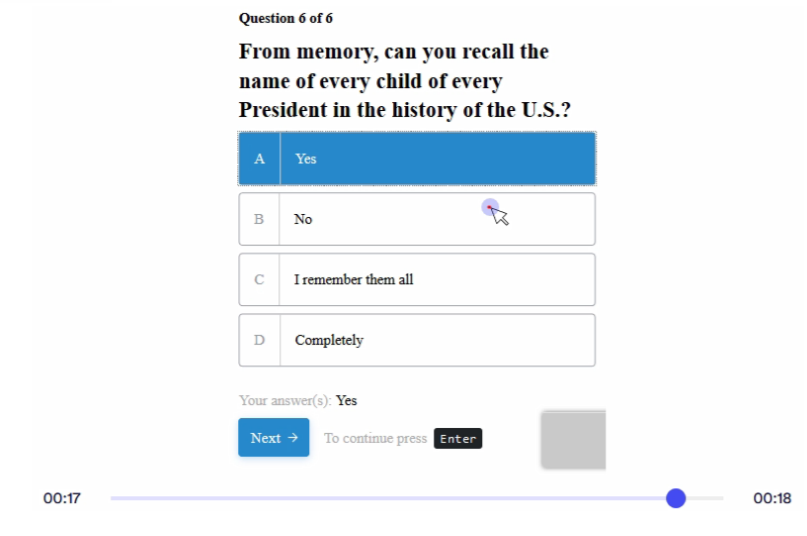












Classification Protocols

Presurvey Screening

If any are flagged:

- Semantic Network Model Stimulus 1
- Semantic Network Model Stimulus 2
- Semantic Network Model Stimulus 3
- Semantic Network Model Stimulus 4
- Yea-Saying 1
- Yea-Saying 2

Identified as Problematic: 28%

Postsurvey Review

If any are flagged:

- Major open-ended response issue
- Easy in-grid trap

If at least two are flagged:

- Bottom 10% LOI
- Minor open-ended response issue
- Moderate in-grid trap
- Straightlining
- All-checking
- Number box issue 1
- Number box issue 2
- Rare event 1
- Rare event 2

Identified as Problematic: 27%

Estimating True Status

True Status Estimate

Classified as problematic if evidence is strong:

- Flagged by both protocols
- Total issues in identifying protocol are > μ+2σ
- Has an issue the protocol developer considers a strong signal of quality status (one allowed per protocol)
  - Presurvey: Fails both yea-saying tasks
  - Postsurvey: High-severity open-ended issue (profanity, nonsense, "gooding")

Identified as Problematic: 21%

Postsurvey Review → Est. True Status

Presurvey Screens	Flag	Pass
Flag	339 ← *35* | 215	
Pass	73 | *162* → 1270	

Likely Problematic

- Failed Both 16%
- Presurvey Alone, Strong 2%
- Postsurvey Alone, Strong 3%

Likely Valid

- Passed Both 61%
- Presurvey Alone, Moderate 10%
- Postsurvey Alone, Moderate 8%

Presurvey Screening Performance

Presurvey Screens	Estimated True Status: Flag	Pass	Total
Flag	374	215	589
Pass	73	1432	1505
Total	447	1647	2094

- Presurvey screens quarantine 84% of problematic respondents
- Using presurvey screens alone, 95% of the sample is valid
- False positive rate is nearly equal to full postsurvey review

	Presurvey	Postsurvey
Accuracy	86%	91%
F1 Score	72%	81%
Cohen's Kappa	0.63	0.75
True Positive	84%	92%
True Negative	87%	90%
False Positive	13%	10%
False Negative	16%	8%

Adjusted Tolerance

- Allowing one failure in presurvey screening lowers the false positive rate to 2%
- Trade-off with true positive (hit) rate (54%)
- 89% of sample would be valid
- Full postsurvey review finds 100% of false negatives
  - In-grid trap alone finds 39%
  - Adding major OE flags finds an additional 45%

Estimated True Status

Presurvey Screens *One Forgiven*

	Flag	Pass	Total
Flag	243	31	274
Pass	204	1616	1820
Total	447	1647	2094

- Accuracy 89%
- Cohen's Kappa 0.61
- F1 Score 67%

- True Positive 54%
- True Negative 98%
- False Positive 2%
- False Negative 46%

Summary

Performance of presurvey screening in classifying problematic respondents:

- Quarantines 84% of problematic respondents
- 13% false positive rate, consistent with full postsurvey review
- 95% of resulting sample is valid

Appendix: Additional Studies

Study 2: Semantic Network Tasks Alone

- Early performance test of semantic network-based stimuli without yea-saying tasks
- Directly comparing presurvey screening with postsurvey review; no estimate of true status (presurvey is too limited to create a balanced true status protocol)

Presurvey Screening

If any are flagged:

- Semantic Network Model Stimulus 1
- Semantic Network Model Stimulus 2
- Semantic Network Model Stimulus 3
- Semantic Network Model Stimulus 4

Identified as Problematic: 19%

Postsurvey Review

If any are flagged:

- Major open-ended response issue
- Screener trap
- In-grid trap

If at least two are flagged:

- Bottom 10% LOI
- Minor open-ended response issue
- Straightlining
- All-checking
- Number box issue
- Rare event 1
- Rare event 2
- Rare event 3

Identified as Problematic: 31%

Study 2: Presurvey Screening Performance

Presurvey Screens	Postsurvey Review: Flag	Postsurvey Review: Pass	Postsurvey Review: Total
Flag	178	82	260
Pass	254	886	1140
Total	432	968	1400

	Presurvey
Accuracy	76%
F1 Score	51%
Cohen's Kappa	0.37
True Positive	41%
True Negative	92%
False Positive	8%
False Negative	59%

- Semantic network-based tasks alone identify 41% of postsurvey's problematic respondents
- Good false positive rate, high false negative rate
- Using these four presurvey tasks alone, 78% of the sample would be valid

Study 3: Limited Postsurvey Review

- A performance test in a more challenging environment for quality review
  - Survey methods and instrumentation allow for fewer opportunities for postsurvey flags
  - Lower qualification rate / lower incidence population

Presurvey Screening

If any are flagged:

- Semantic Network Model Stimulus 1
- Semantic Network Model Stimulus 2
- Semantic Network Model Stimulus 3
- Semantic Network Model Stimulus 4
- Yea-Saying 1
- Yea-Saying 2

Identified as Problematic: 35%

Postsurvey Review

If any are flagged:

- Major open-ended response issue
- In-survey attention check

If at least two are flagged:

- Bottom 10% LOI
- Minor open-ended response issue
- Number box issue
- Rare event

Identified as Problematic: 18%

Study 3: Presurvey Screening Performance

Presurvey Screens	Estimated True Status: Flag	Pass	Total
Flag	292	235	527
Pass	51	916	967
Total	343	1151	1494

	Presurvey	Postsurvey
Accuracy	81%	94%
F1 Score	67%	85%
Cohen's Kappa	0.54	0.81
True Positive	85%	76%
True Negative	80%	99%
False Positive	20%	1%
False Negative	15%	24%

- Presurvey screening's hit rate outperforms what a more limited postsurvey review can achieve, identifying 85% of problematic respondents (vs. 76% with postsurvey)
- 95% of the resulting sample is valid (vs. 93% with postsurvey)
- Adjusted tolerance (1 presurvey failure): 2% false positive rate, 69% true positive rate