

GALLUP®

# Monitoring Data Quality in Probability-Based Internet Panels

STEPHEN RAYNES, PH.D.  
JENNY MARLAR, PH.D.

GALLUP



## PRESENTATION AGENDA

- 1 Background and Objectives
- 2 Experimental Methods
- 3 Analysis Results
- 4 Summary and Future Research

# Probability-Based vs. Non-Probability (Opt-In) Online Samples

## Extensive evidence that probability panels tend to provide higher quality data than opt-in samples:

- Kennedy, Mercer & Lau (2024)
- Herman et al. (2024)
- Callegaro, Villar, Yeager, and Krosnick (2014)
- Dutwin & Buskirk (2017)
- MaInnis, Krosnick, Ho, and Cho (2018)
- Lorenz, Marlar, Han & Natzke, Gallup (AAPOR 2023)
- Mercer & Lau, Pew Research (2023)

Probability-based samples	Non-probability (opt-in) samples
Sampled from a known frame (e.g., ABS, RDD)	Respondents self-select into panels or studies
Known, non-zero selection probabilities	Selection probabilities unknown
Recruitment occurs before participation	Participation driven by incentives and availability
Transparent and defensible methodologies	Less transparent methodologies
Higher cost per survey	Lower cost per survey

# Probability-Based Panels Reduce Selection Error, But Measurement Risks Evolve

## PROBABILITY-BASED PANELS

- Most recent “data quality” discussions in probability panels focus on selection and representation (recruitment, attrition, contactability, nonresponse, weighting).
- A potential double-edged sword in modern probability panels: incentives
  - Reduce nonresponse bias
  - May increase risk of careless responding behaviors

### Research Questions:

1. How prevalent are careless-responding signals in modern probability panels?
2. How should we monitor and address them?

We compare the *same flagging framework* to an opt-in experiment we ran (AAPOR 2025) to contextualize the magnitude of careless responding rates



## PRESENTATION AGENDA

- 1 Background and Objectives
- 2 Experimental Methods
- 3 Analysis Results
- 4 Summary and Future Research

# Study Methods

## Panel Survey

- Web only, English only
- Fielded Fall, 2025
- U.S Adults
- Median length: 6.8 minutes
- Sample size: 2220
- No exclusions applied

## Comparison: Opt-in Experiment Survey

- Web only, English only
- Fielded Spring, 2025
- U.S Adults
- Median length: 8.5 minutes
- Sample size: 7069 (total across 7 vendors)
- No exclusions applied
- Slightly different survey (e.g. attention checks, bogus items included)

## (Mostly) Overlapping Survey topics included:

- Health and wellbeing
- Home ownership/renting
- Employment status
- Civic engagement
- Demographics

Results shown are unweighted unless noted

# 21 Unique Careless Responding Flags Assessed on a Pass/Fail Basis

## SURVEY-EMBEDDED ITEMS (1)

### Qualitative items

- Open-end gibberish

## POST-HOC INDEXES (15)

### Response time

- Total duration speeding
- Total duration slowing
- Page-level speeding
- Survey acceleration

### Suspicious Response Patterns

- Partial survey completes
- Acquiescence bias
- DK/PNA bias

### Response consistency

- Logical consistency
- Repeat HH size consistency
- *Psychometric synonyms\**
- *Psychometric antonyms\**

### Individual Variability

- Maximum longstring straightlining
- Average longstring straightlining
- Individual response variability (IRV)

### Outliers

- *Mahalanobis distance\**

## SUSPICIOUS ACTIVITIES (5)

### Exact Duplications

- Across key batteries
- Unique participant ID
- Duration or finish time

### Majority Duplication

- Entire survey

### Automated

- Qualtrics ReCaptcha score

*\*These diagnostics were included for comparability with opt-in results, but the survey design did not meet the assumptions required for them to function as valid psychometric or multivariate outlier tests.*



## PRESENTATION AGENDA

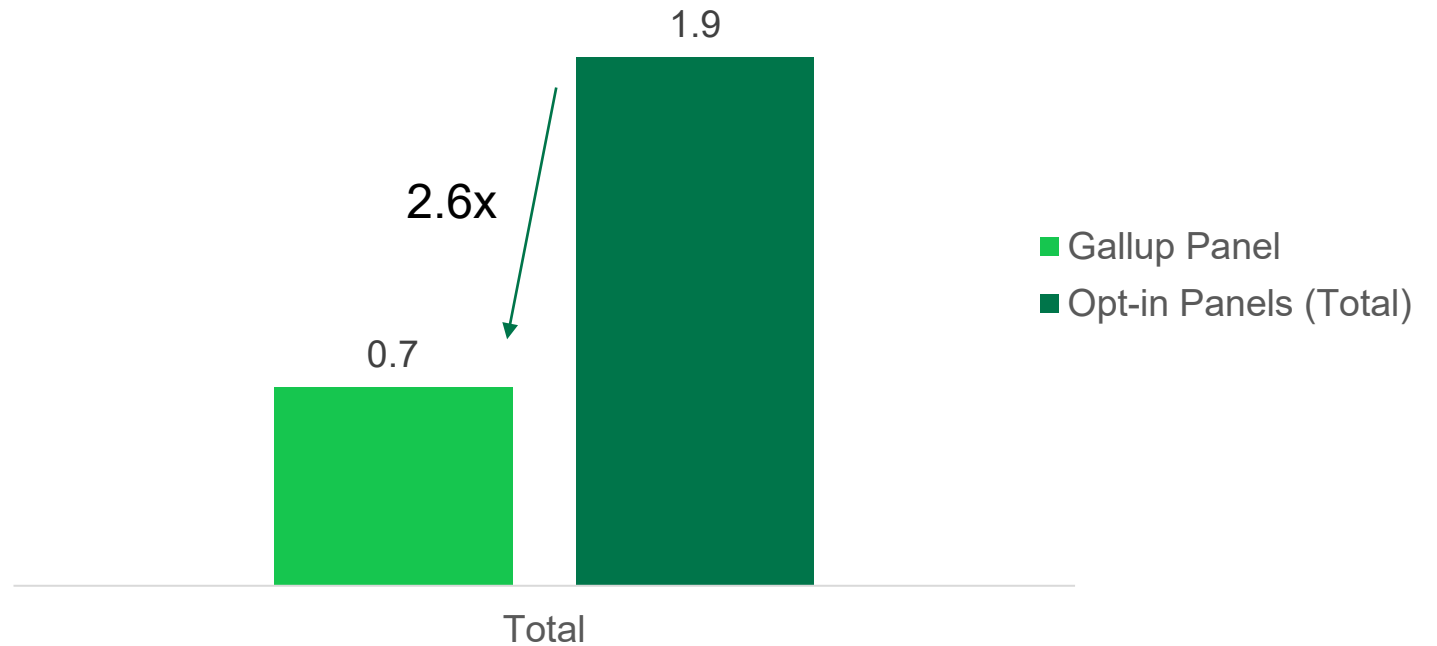
- 1 Background and Objectives
- 2 Experimental Methods
- 3 Analysis Results
- 4 Summary and Future Research

# Probability Panel Respondents Fail 2.6x Fewer Flags On Average than Opt-in

## When considering the total average number of flags per person:

- Mean flags failed is  $<1$ , but not 0
- Mean flags failed is **2.6x lower** than opt-in

Mean Total Flags Per Person



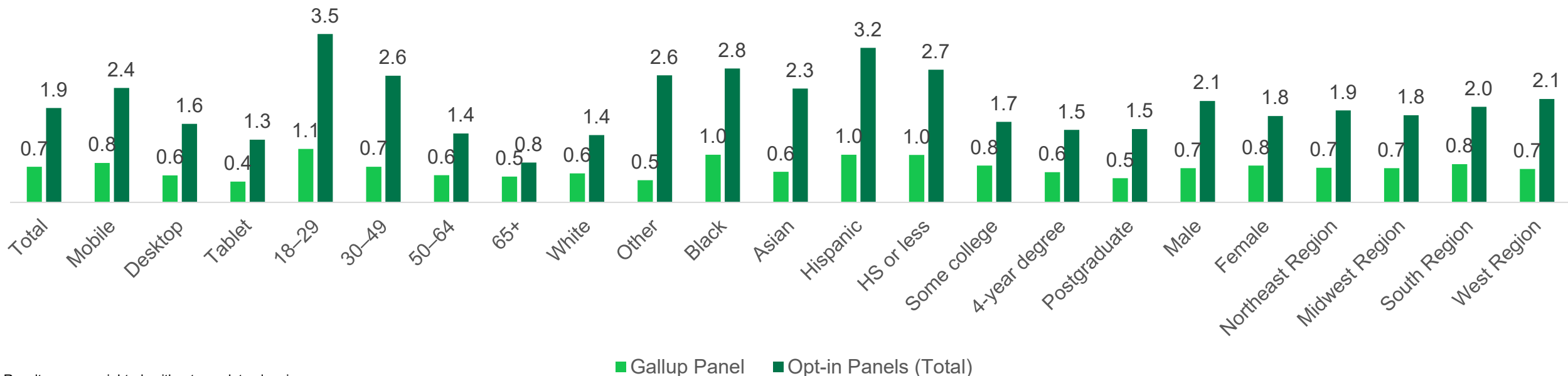
Results are unweighted, without any data cleaning

# Average Number of Flags Failed Vary Modestly Across Demographics

## When considering the total average number of flags per person:

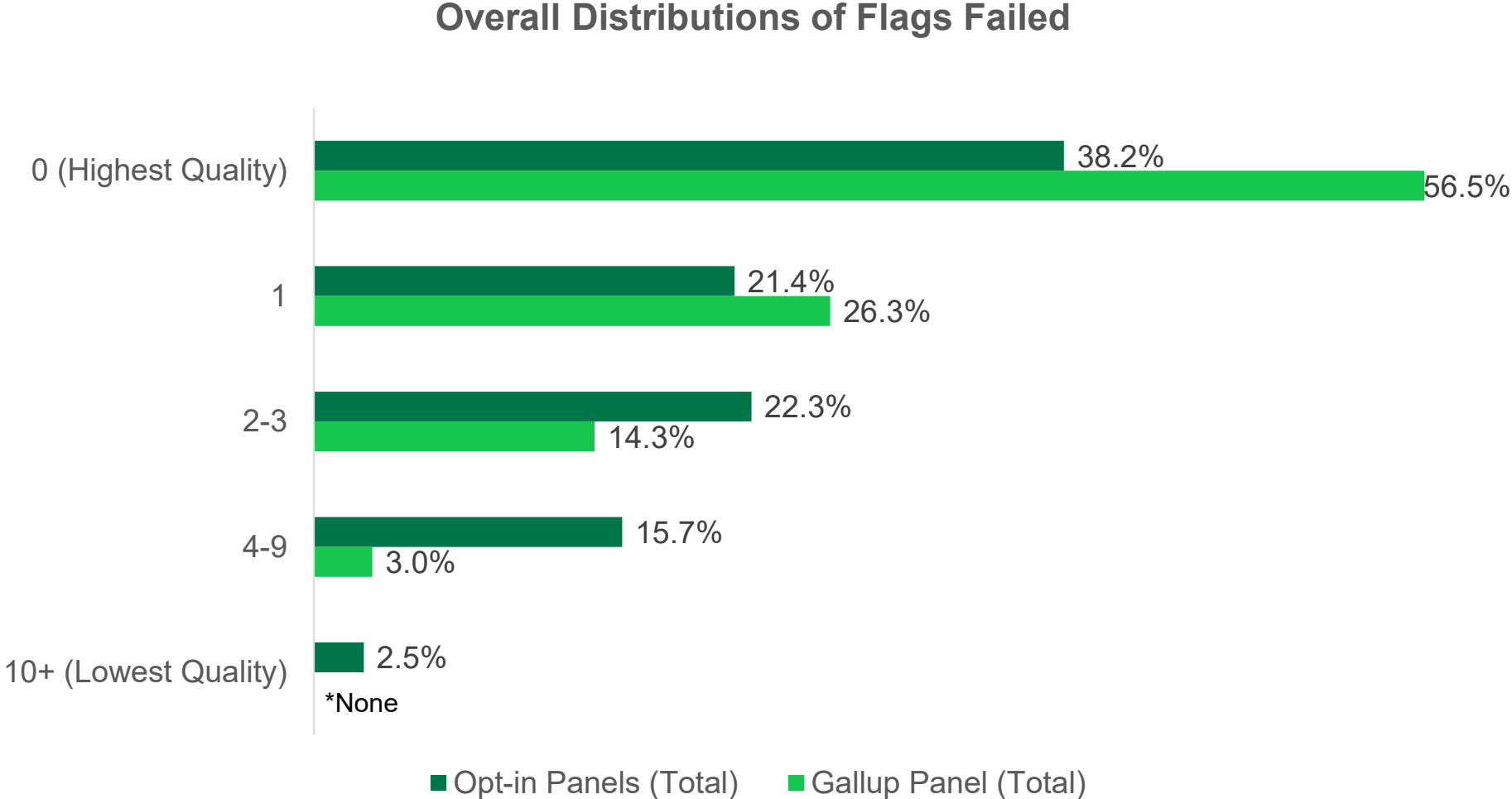
- Fail rates vary by subgroup in familiar directions (e.g., younger, lower education), however:
  - Some careless responding is detected in every subgroup – not isolated to specific populations
  - There is significantly less variance across subgroups in probability panel compared to opt-in panel results

Mean Total Flags Per Person



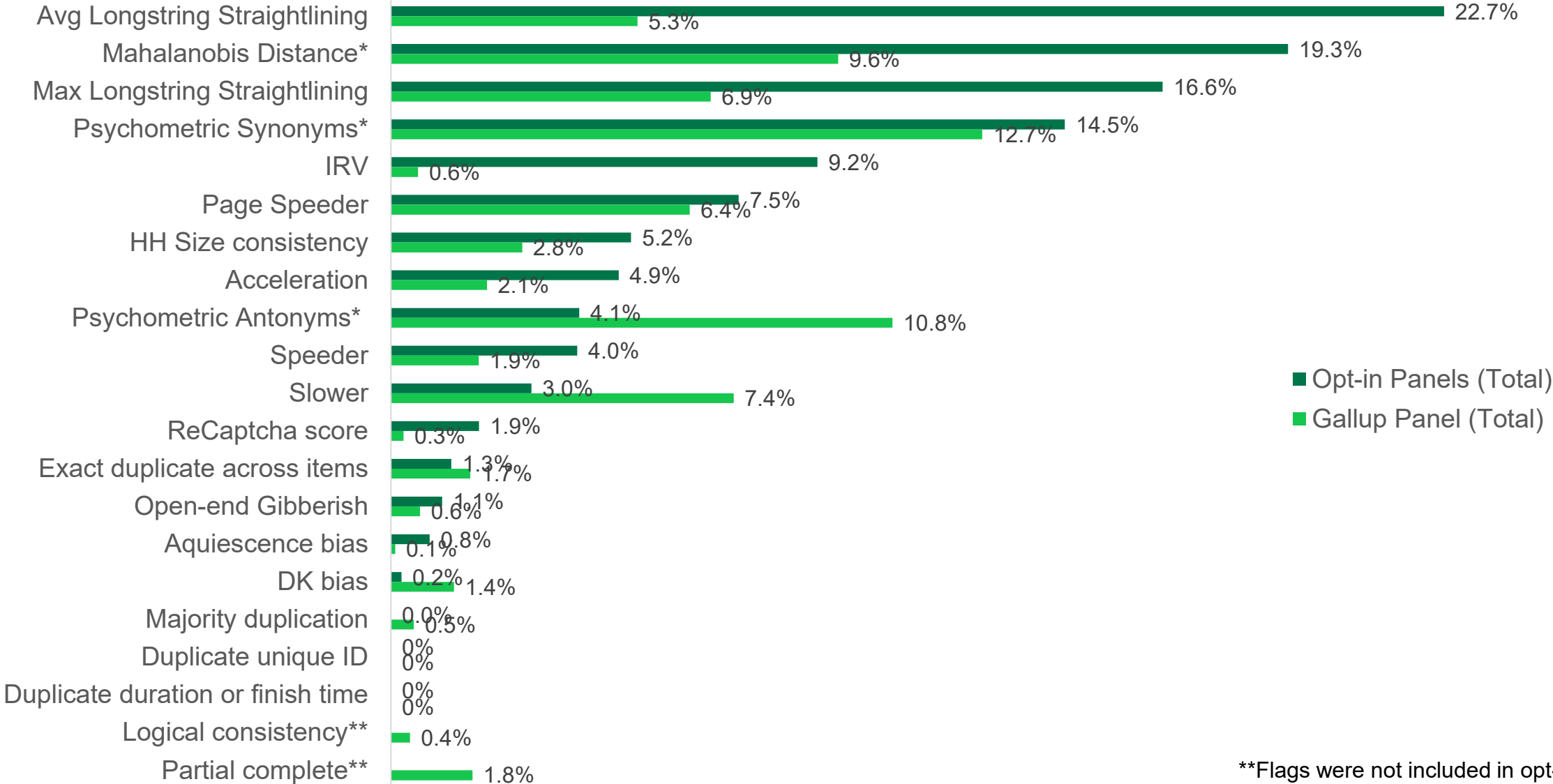
Results are unweighted, without any data cleaning

# Over Half of Probability Panelists Pass All 21 Possible Data Quality Flags



# Opt-in Panelists More Frequently Fail Almost All Possible Flags

Frequency of Individual Flags Hit



\*\*Flags were not included in opt-in experiment

\*These diagnostics were included for comparability with opt-in results, but the survey design did not meet the structural assumptions required for them to function as valid psychometric or multivariate outlier tests.

# Which Data Quality Flags Should Probability-Based Panels Routinely Utilize?

Recommendation: start with a small set of scalable, interpretable flags based on duration, consistency, and automated suspicious activity checks

<b>Data Quality Flag</b>	<b>Fail Rate: % Failed Overall</b>	<b>Uniqueness of Flag: % Failed This Flag and No Other Flags</b>	<b>Comprehensiveness of Flag: Avg. # of Additional Flags Failed Among Those Failing This Flag</b>
Majority duplication	0.5%	0.0%	5.7
Exact duplicate across items	1.7%	0.0%	4.4
Speeder	1.9%	0.0%	3.9
Partial complete	1.8%	0.1%	3.3
Acceleration	2.1%	0.3%	3.0
Max Longstring Straightlining	6.9%	0.6%	2.2
Avg Longstring Straightlining	5.3%	0.6%	1.7
DK bias	1.4%	0.3%	1.6
Page Speeder	6.4%	2.6%	1.4
HH Size consistency	2.8%	0.8%	1.1
IRV	0.6%	0.2%	1.0
Mahalanobis Distance*	9.6%	4.3%	0.9
Slower	7.4%	4.1%	0.8
Psychometric Synonyms*	12.7%	6.2%	0.8
Psychometric Antonyms*	10.8%	5.7%	0.7
ReCaptcha score	0.3%	0.1%	0.5
Logical consistency	0.4%	0.1%	0.4
Open-end Gibberish	0.6%	-	0.3
Acquiescence bias	0.1%	0.0%	0.1
Duplicate unique ID	0.0%	0.0%	-
Duplicate duration or finish time	0.0%	0.0%	-

*\*These diagnostics were included for comparability with opt-in results, but the survey design did not meet the structural assumptions required for them to function as valid psychometric or multivariate outlier tests.*

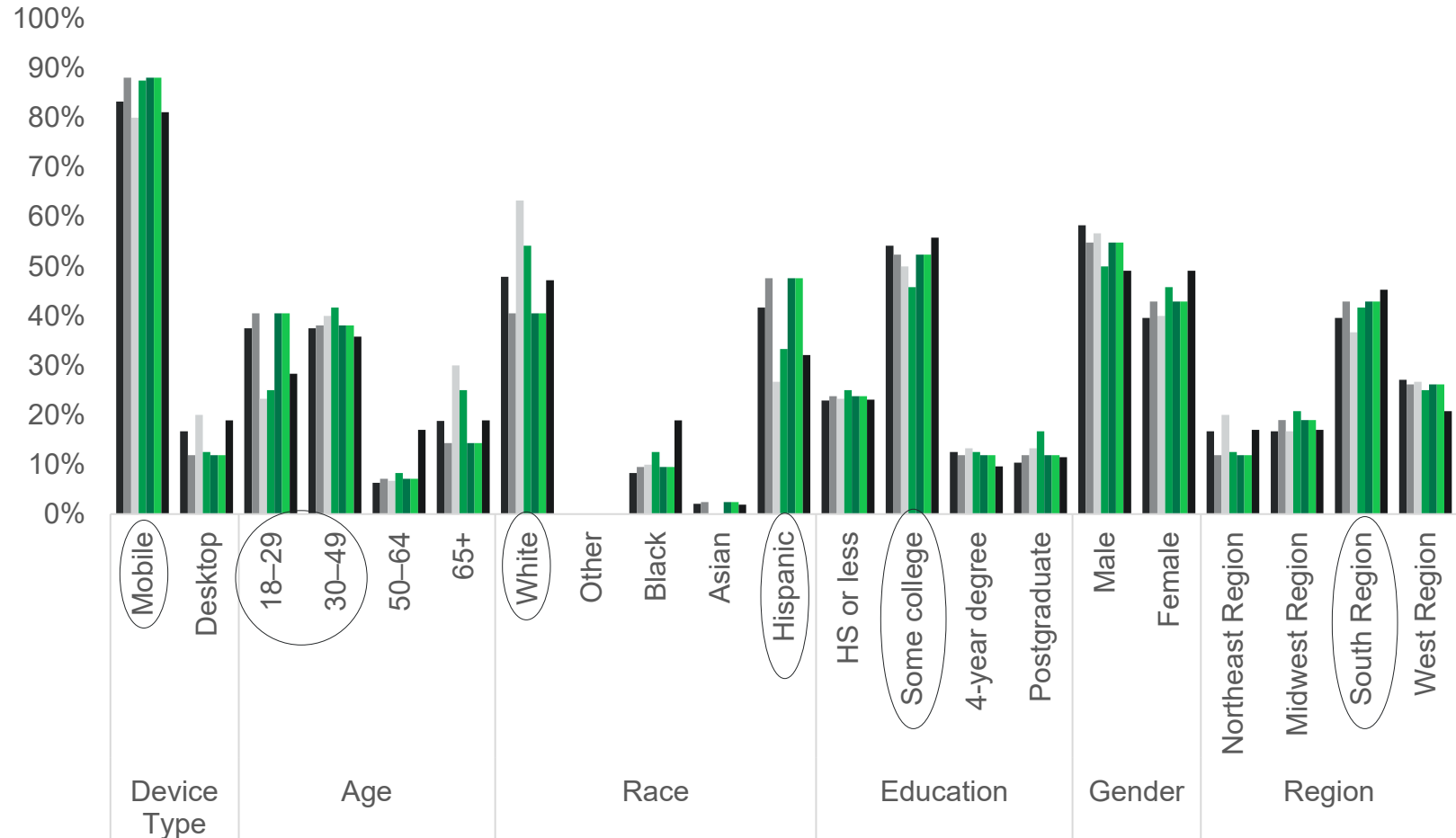
# About 2% of Respondents Flagged for Possible Data Quality Issues Across 7 Flagging Approaches

Demographic Distributions Are Similar Across Strategies, and Far Less Extreme Than in Opt-In Samples

## 7 Flagging Approaches:

- ReCaptcha or Speeder (N=48)
- Speeder (N=42)
- ReCaptcha or (speeder AND straightliner) (N=30)
- Speeder and straightliner (N=24)
- Speeder or (any flag for lack of variability) (N=42)
- Speeder or (straightliner and ReCaptcha and inconsistent response) (N=42)
- 2+ of any of 3 key careless flags: speeder, straightliner, or inconsistent responses; OR 2+ of any of the 5 "fraud-like" flags (N=53)

### Demographic Distributions of Each Flagging Approach





## PRESENTATION AGENDA

- 1 Background and Objectives
- 2 Experimental Methods
- 3 Analysis Results
- 4 Summary and Future Research

# Summary of Research Findings and Key Recommendations

01

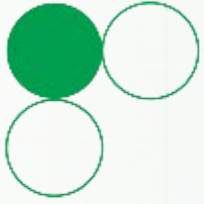
## How prevalent are careless-responding signals in modern probability panels?

- Careless-responding signals are extremely low (~2% across methods) and relatively randomly distributed compared to opt-in
- While current careless responding patterns are not alarming, ongoing monitoring of measurement error is increasingly necessary to maintain high-quality probability-based panels.

02

## How should we monitor and address them?

- There is no single silver-bullet detection or removal strategy for careless responders.
- A practical approach:
  - Monitor a small, interpretable core set of flags
  - Use flagged cases for sensitivity analyses before exclusion on individual surveys
  - Track flagged cases across individual surveys to identify consistent data quality offenders in your panel



Thank you!

Stephen Raynes  
[stephen\\_raynes@gallup.com](mailto:stephen_raynes@gallup.com)

Jenny Marlar  
[jenny\\_marlar@gallup.com](mailto:jenny_marlar@gallup.com)



---

# Appendix

# Flag Definitions

variable_name	description
flag_careless_OpenEnd	Fed open-end response to "What is the most important thing that could make your life better in the next five years?" into ChatGPT 5.0 to flag clearly invalid responses (excluding skips). Prompt documented in code comments.
flag_careless_partial	Partial complete (PARTICIPATION_ID == 10)
flag_careless_speeders	< 1/3rd median total study duration
flag_careless_slowers	> 10x median total study duration
flag_careless_speeders_page	Submitted >50% of pages too fast (faster than 3 seconds per item × items per page)
flag_careless_acceleration	Survey split into thirds; flagged if 2nd third is ≥2× faster than 1st AND final third ≥2× faster than 2nd
flag_careless_aquiescence	Selected top affirmatory/positive pole for ≥2/3 of ordinal items (32 of 48)
flag_careless_dk	Selected Don't know / Prefer not to respond / skipped ≥2/3 of items
flag_careless_HH_consistency	Household size inconsistency (ALIGNMENT_1 vs ALIGNMENT_2)
flag_careless_inconsistent_poli	Political ideology inconsistency (e.g., Republican & very liberal; Democrat & very conservative)
flag_careless_straight_maxlongstring	Outlier max longstring on GPSS and MEDIA question sets
flag_careless_straight_avglongstring	Outlier average longstring on GPSS and MEDIA question sets
flag_careless_irv	Outlier IRV on GPSS and MEDIA question sets
flag_careless_psychsyn	Insufficient positive correlation on psychometric synonym pair (MEDIA set)
flag_careless_psychant	Insufficient negative correlation on psychometric antonym pair (GPSS set)
flag_careless_mahal	Mahalanobis distance outlier using 99% CI threshold (GPSS or MEDIA)
flag_bot_exactdup	Exact duplication across all GPSS, PROBLEM, and MEDIA items
flag_bot_dupID	Duplicate unique respondent ID
flag_bot_duplication70	>70% duplication across all survey items answered
flag_bot_duptime	Duplicate exact study duration or identical EndDate
flag_bot_reCaptcha	< 0.5 Qualtrics reCAPTCHA score

# Citations

**Callegaro, M., Villar, A., Yeager, D., & Krosnick, J. A. (2014).**

A critical review of studies investigating the quality of data obtained with online panels based on probability and nonprobability samples. In M. Callegaro, R. Baker, J. Bethlehem, A. S. Göritz, J. A. Krosnick, & P. J. Lavrakas (Eds.), *Online panel research* (pp. 23–53). Wiley.

<https://doi.org/10.1002/9781118763520.ch2>

**Dutwin, D., & Buskirk, T. D. (2017).**

Apples to oranges or Gala versus Golden Delicious? Comparing data quality of nonprobability Internet samples to low response rate probability samples. *Public Opinion Quarterly*, 81(S1), 213–239.

<https://doi.org/10.1093/poq/nfw061>

**Herman, P. M., et al. (2024).**

Comparing health survey data cost and quality between Amazon’s Mechanical Turk and Ipsos’ KnowledgePanel: Observational study. *Journal of Medical Internet Research*, 26, e63032.

<https://www.jmir.org/2024/1/e63032>

**Kennedy, C., Mercer, A., & Lau, A. (2024).**

Exploring the assumption that commercial online nonprobability survey respondents are answering in good faith. *Survey Methodology*, 50(1). Statistics Canada.

<https://www150.statcan.gc.ca/n1/pub/12-001-x/2024001/article/00013-eng.htm>

**Lorenz, J., Marlar, J., Han, S., & Natzke, J. (2023).**

Practical considerations for mixed-panel research using probability-based and opt-in samples. Paper presented at the *Annual Conference of the American Association for Public Opinion Research (AAPOR)*, Philadelphia, PA.

**MacInnis, B., Krosnick, J. A., Ho, A. S., & Cho, M.-J. (2018).**

The accuracy of measurements with probability and nonprobability survey samples: Replication and extension. *Public Opinion Quarterly*, 82(4), 707–744.

<https://doi.org/10.1093/poq/nfy038>

**Mercer, A., & Lau, A. (2023).**

Comparing two types of online survey samples: Opt-in samples are about half as accurate as probability-based panels. *Pew Research Center (Methods)*.

<https://www.pewresearch.org/methods/2023/09/07/comparing-two-types-of-online-survey-samples/>