Coming Clean: Does Data Cleaning Reduce or Increase Bias in Sub-groups?

Randall K. Thomas, Frances M. Barlas,

and Megan A. Hendrich

Ipsos Public Affairs

FCSM Federal Committee on Statistical Methodology

2021 FCSM Research & Policy Conference
November 2nd – November 4th, 2021

Background

- Many researchers believe that it is necessary to clean survey data before analysis in order to improve accuracy.
- One specific concern has been about sub-optimal, or less-than-accurate, response.
- Sub-optimal response is seen as a source of lower quality data:
    - a dishonest or mistaken response (a bias)
    - an inattentive response (error), or
    - an approximate response rather than the respondent's true answer (some good measurement plus some error)

Background

- Researchers have developed many measures of sub-optimal response, including :
    - Speeding through the survey
    - Grid non-differentiation or straightlining
    - Item nonresponse – skipping items
    - Extreme or exaggerated responding on numeric entry
    - Failure at trap questions (e.g., compliance traps)
    - Consistency checks

Research Questions

- While there is little research on data cleaning and suboptimal response, what does exist seems to start with the assumption that data cleaning is necessary to improve the accuracy of survey results.
- With Study 1, we set out to test that assumption and sought to address:
    - How much data cleaning is necessary?
    - How should we clean our data?
    - Is it possible to do too much data cleaning?
- In Study 2, we sought to examine how cleaning affects survey results for specific subgroups.

Study 1 - Method

Study 1 Method

In October 2015, we conducted parallel studies using two online sample sources:

Ipsos KnowledgePanel

- The largest probability-based panel in the U.S. Sample obtained primarily through ABS recruitment.
- Obtained 1,297 completes

Non-probability sample – opt-in sample

- To obtain a demographically balanced sample, we set up interlocking quotas by gender, age, race/ethnicity, and education
- Obtained 2,564 completes

Sample Cleaning Criteria and Size

Excluded (%)	KnowledgePanel		Non-probability Sample	
	Size	Cleaning Criteria	Size	Cleaning Criteria
0	1,297	None	2,564	None
2.5	1,265	Item NR, Speed	2,497	Item NR, Speed
5	1,232	Item NR, Speed, Grids	2,431	Item NR, Speed, Grids
10	1,172	Item NR, Speed, Grids, Numeric	2,304	Item NR, Speed, Grids, Numeric
20	1,029	Item NR, Speed, Grids, Numeric	2,039	Item NR, Speed, Grids, Numeric
30	909	Item NR, Speed, Grids, Numeric	1,794	Item NR, Speed, Grids, Numeric
40	784	Item NR, Speed, Grids, Numeric	1,537	Item NR, Speed, Grids, Numeric
50	650	Item NR, Speed, Grids, Numeric	1,281	Item NR, Speed, Grids, Numeric

After each round of cleaning, remaining cases were weighted to Current Population Survey demographic benchmarks.

Benchmarks for Bias Evaluation

American Community Survey (2014)

- Home with 2 or fewer bedrooms
- Own 2 or more vehicles
- Married
- Household size 2+
- Employed
- Own home

CPS–Civic Engagement (2012)

- Discuss politics with family and friends
- Voting in local elections
- Contact with friends/family
- Trust people in neighborhood

National Health Interview Survey (2013)

- Working landline
- General health – good or better
- More than 1 year since doctor visit
- Doctor not taking new patients
- Lifetime drinker
- Current smoker
- Sleep 7+ hours

CPS – Volunteer Supplement (2014)

- Volunteer in last year
- Donate $25 or more

CPS – Food Security (2014)

- Need to spend more on food

General Social Survey (2014)

- Favor death penalty
- Most people can be trusted
- Women less likely to be promoted
- Religiosity

CPS – Public Participation in the Arts (2012)

- Visited art museum or gallery
- Visited park or monument
- Read a book in last year
- Went to the movies
- Went to sporting event
- Worked with plants/ gardened

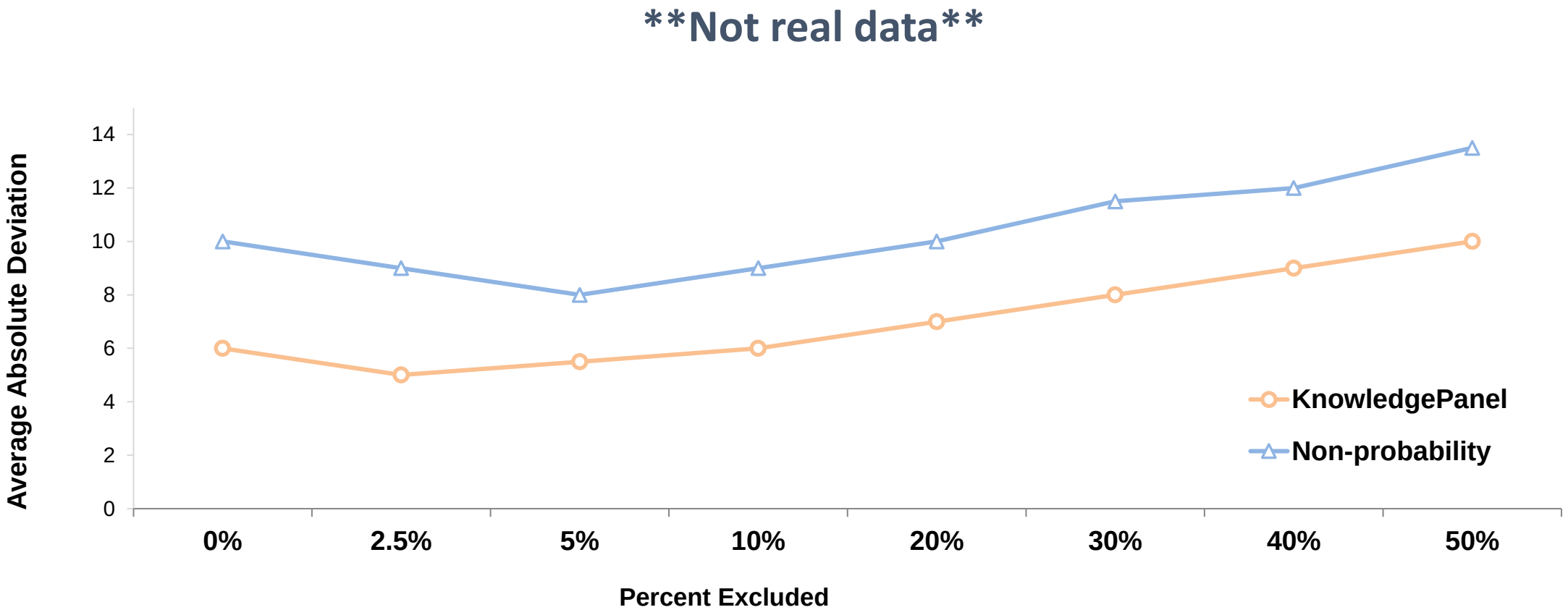
Average Absolute Deviation

To assess bias among the cleaned subsets – calculated the average absolute deviation:

- **Calculated difference between the benchmark and the estimate from each cleaned subset**
- **Took the absolute value for each difference**
- **Averaged across absolute values for each benchmark to obtain average absolute deviation with each cleaned subset**

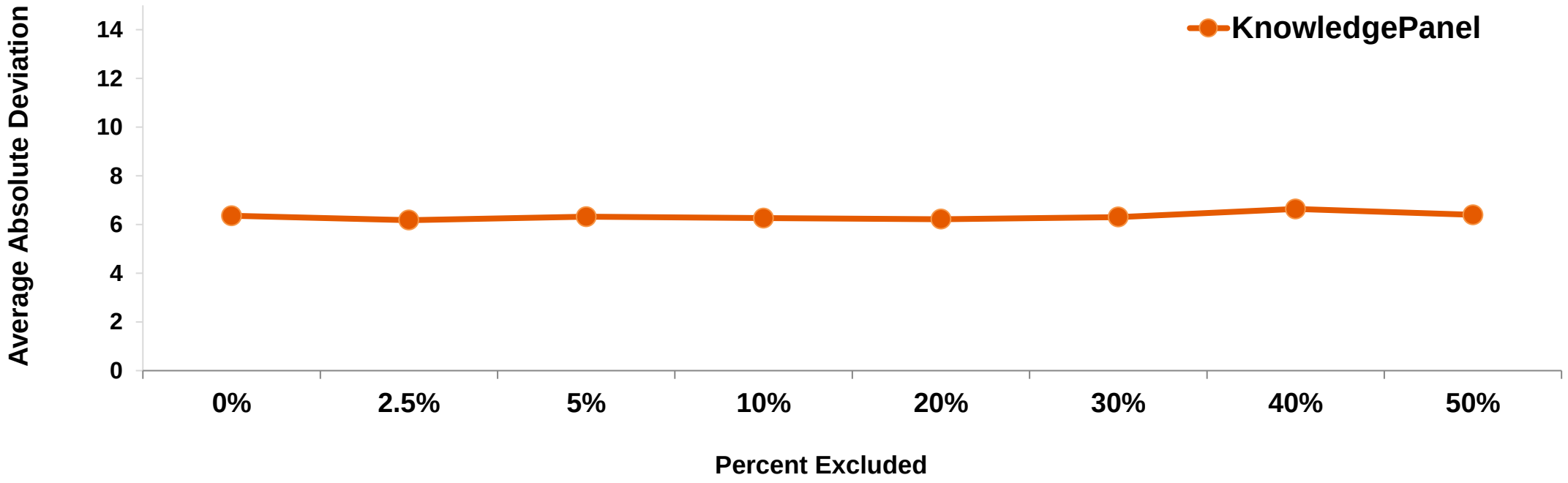
Study 1 Results

Hypothetical Distribution



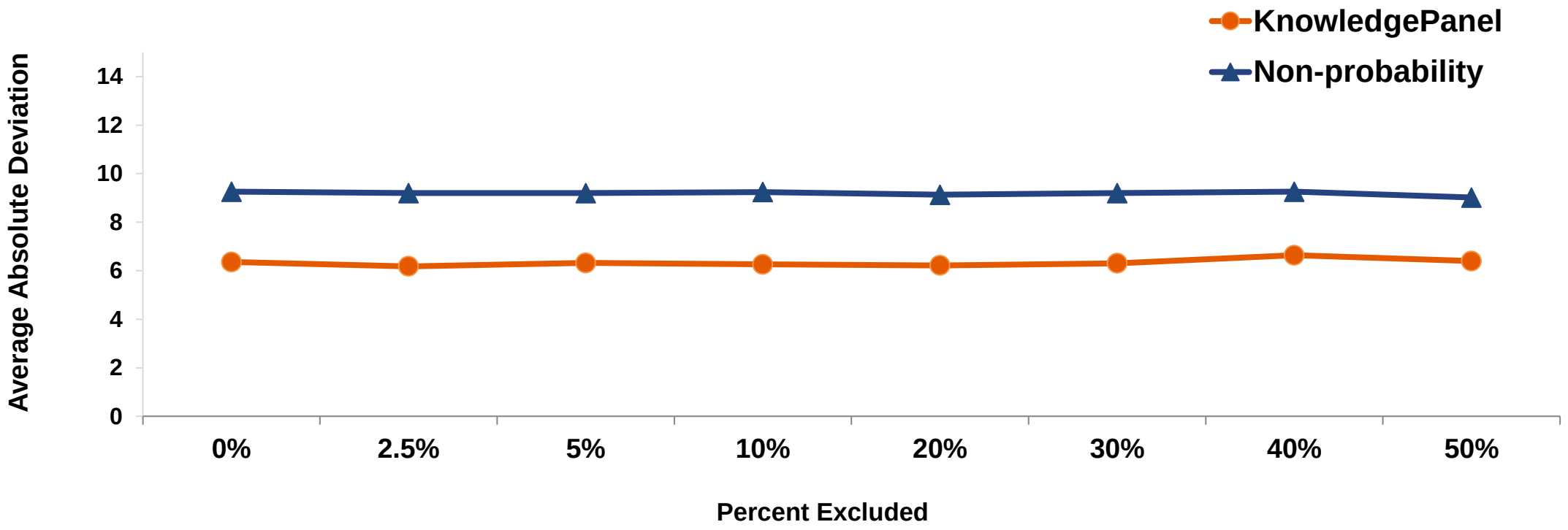
We initially hypothesized that minimal data cleaning, around 5% to 10%, would reduce bias, but extensive cleaning would do more harm than good.

Study 1 Results – Multiple Cleaning Criteria



With KnowledgePanel, no effect on bias with increasingly rigorous exclusion criteria.

Study 1 Results – Multiple Cleaning Criteria



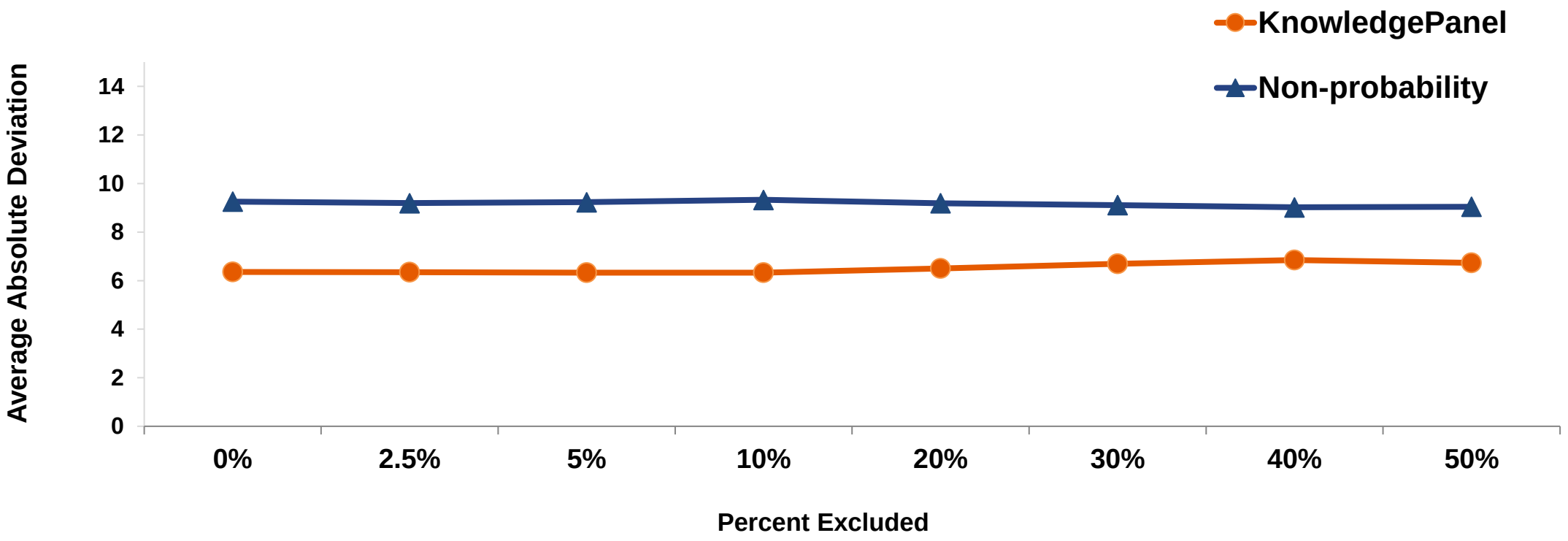
Similarly with non-probability sample, no effect on bias with increasingly rigorous exclusion criteria.

Study 1 Sample Cleaning Criteria and Size

Next, we used time to complete the survey ('speeders') as the sole criterion to use to clean the data. As before, after each round of cleaning, remaining cases were weighted to Current Population Survey demographic benchmarks.

Excluded (%)	KnowledgePanel		Non-probability Sample	
	Size	Cleaning Criteria	Size	Cleaning Criteria
0	1,297	None	2,564	None
2.5	1,264	Speed	2,499	Speed
5	1,231	Speed	2,434	Speed
10	1,163	Speed	2,307	Speed
20	1,029	Speed	2,050	Speed
30	901	Speed	1,794	Speed
40	777	Speed	1,535	Speed
50	647	Speed	1,282	Speed

Study 1 Replicated Excluding Fastest Completes



Rather than using multiple criteria to clean, using speed to complete as the consistent cleaning criterion, we again found no effect on bias with increasingly rigorous exclusion.

Study 2 Research Questions

- **Surprisingly, we found no real change in the overall results or reduction of bias as more and more sample was deleted for both sample types. This was true using both multiple criteria or a single criterion of speed to complete.**
- **One issue that we have noticed in other studies, sometimes deletions due to suboptimal response or due to speed sometimes appeared to impact some groups more that proportionally smaller – including those who are younger, more male, and more people of color.**
- **Our interest in Study 2 was to compare how cleaning might affect the bias we find for specific, smaller groups – White, Black, and Latino.**

Study 2 - Method

Study 2 Methods

In October 2020, we conducted parallel studies using two online sample sources:

- **Ipsos KnowledgePanel (KP) – N=3,344**
- **Non-probability sample – 2 types:**
  - **Opt-in with demographic quotas – using quotas to more adequately demographically balanced the sample, with gender, age, race/ethnicity, and education quotas (N=2,677)**
  - **Opt-in with no demographic quotas (N=3,293)**

Sample 2 Cleaning Criteria and Size

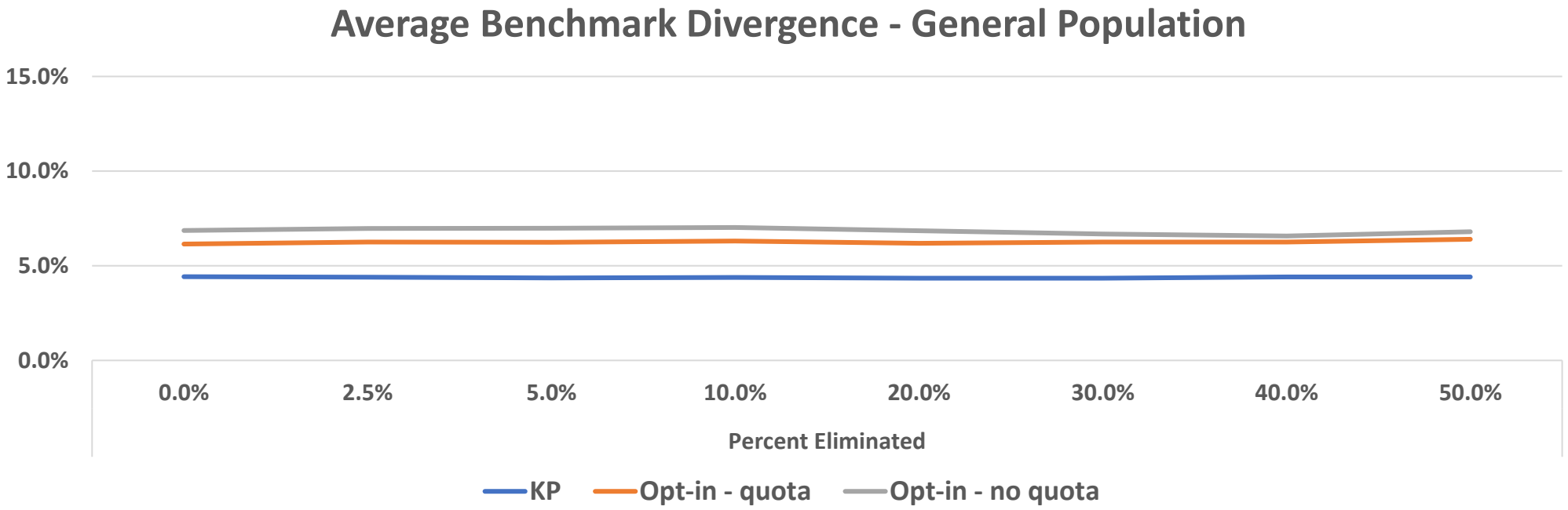
For this analysis, we used speed of completion as the primary criterion for cleaning and created groups within each sample type that eliminated 0%, then the fastest 2.5%, 5.0%, 10.0%, 20.0%, 30.0%, 40.0%, and 50.0%.

After each round of cleaning, for each sample source, remaining cases were weighted to Current Population Survey demographic benchmarks.

We had 25 benchmarks that we were able to obtain national estimates overall for 18+, as well as for White, Black, and Latino groups.

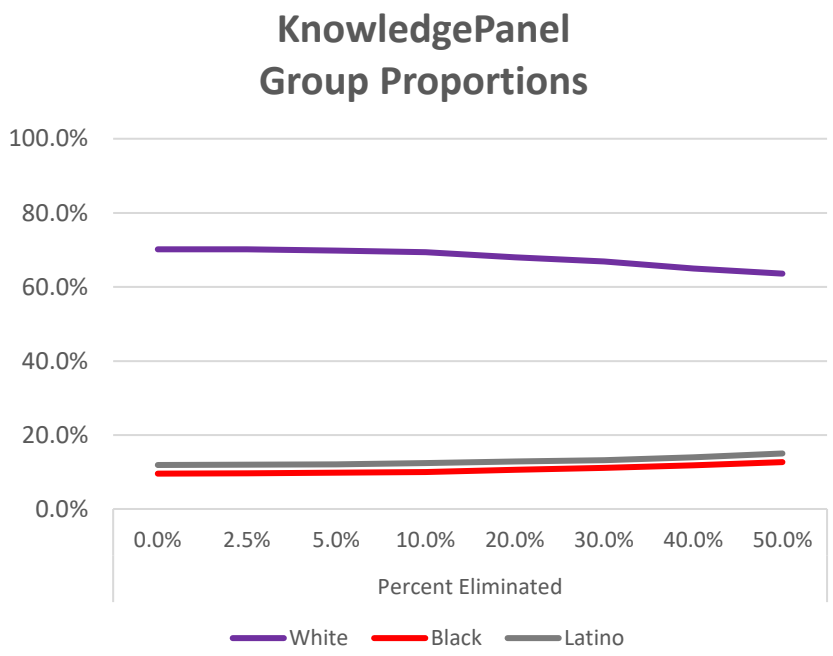
Study 2 Results

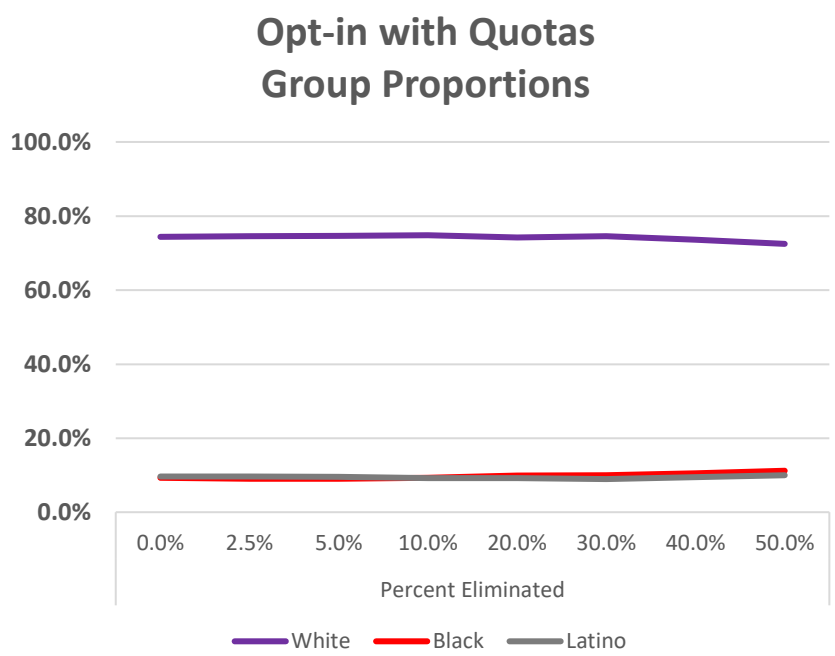
Study 2 Results – Cleaning for Speed and Bias

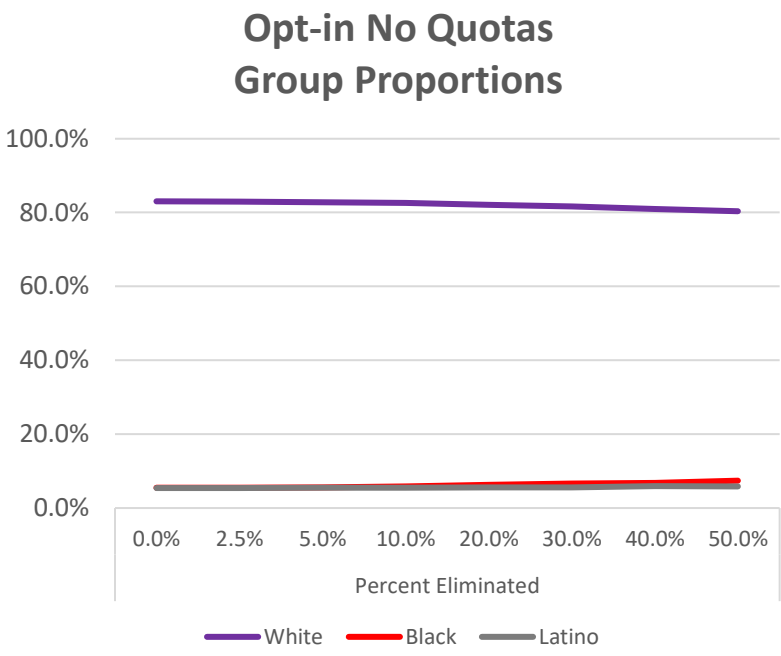


Among all the sample sources, KnowledgePanel again had the lowest bias overall and there was no effect on bias with increasingly rigorous exclusion criteria. Both opt-in samples had higher bias, the no quota opt-in sample showed the highest bias. Increased deselection based on speed may have slightly reduced bias for the no quota opt-in sample.

Study 2 Results – Effects of Cleaning on Group Proportions

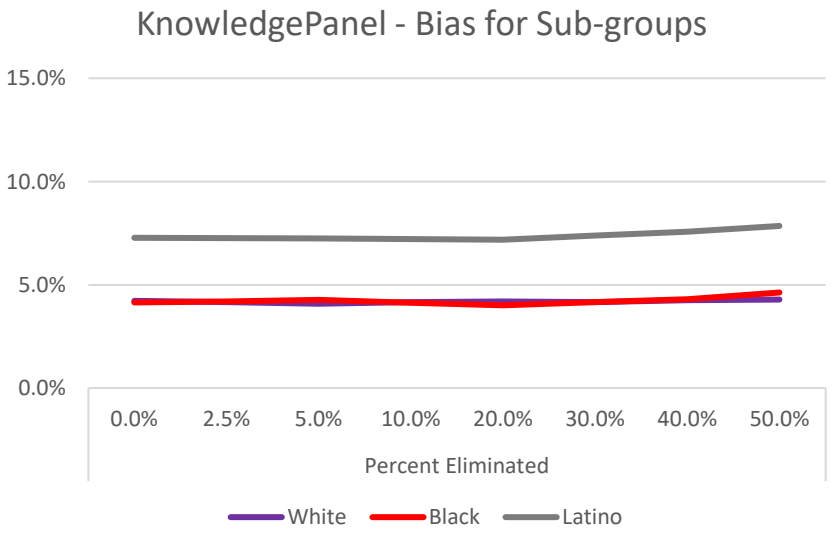


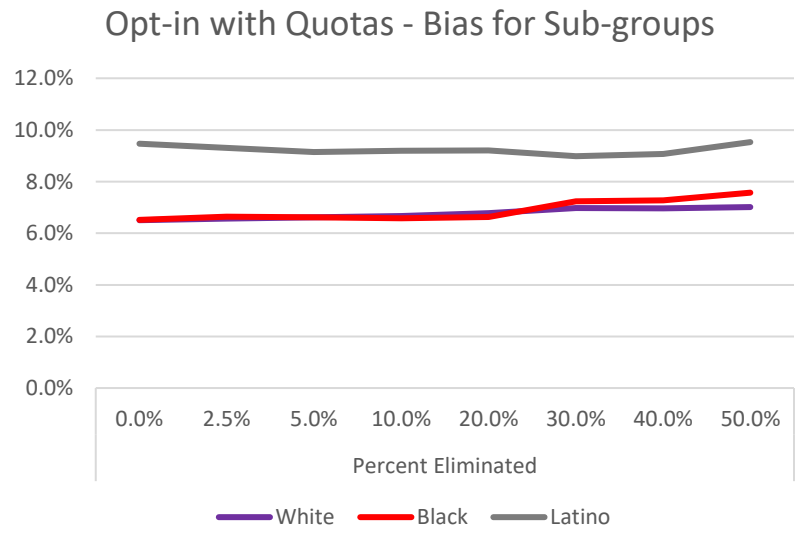


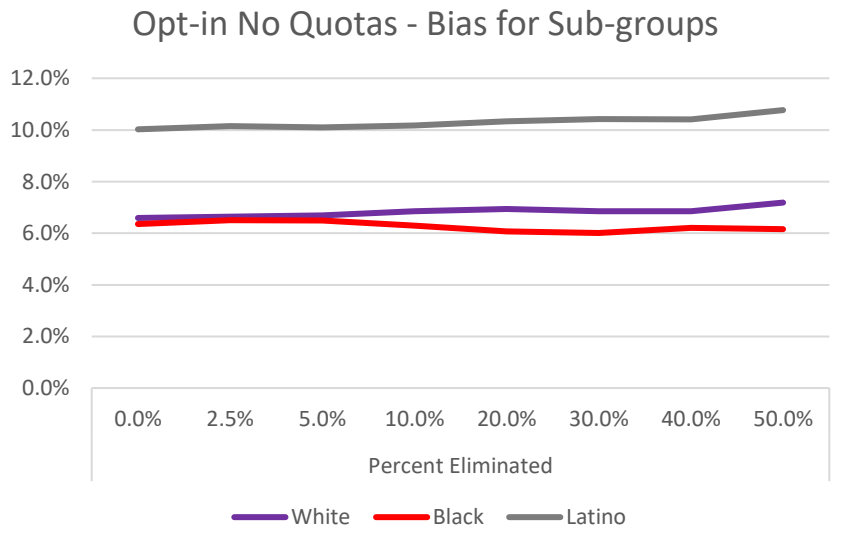


We examined how unweighted group proportions were affected by data cleaning. We found little effect on the relative group proportions – just a very slight increase in Black and Latino under heaviest cleaning for KnowledgePanel and Opt-in with Quotas.

Study 2 Results – Cleaning for Speed and Bias – Race-ethnicity







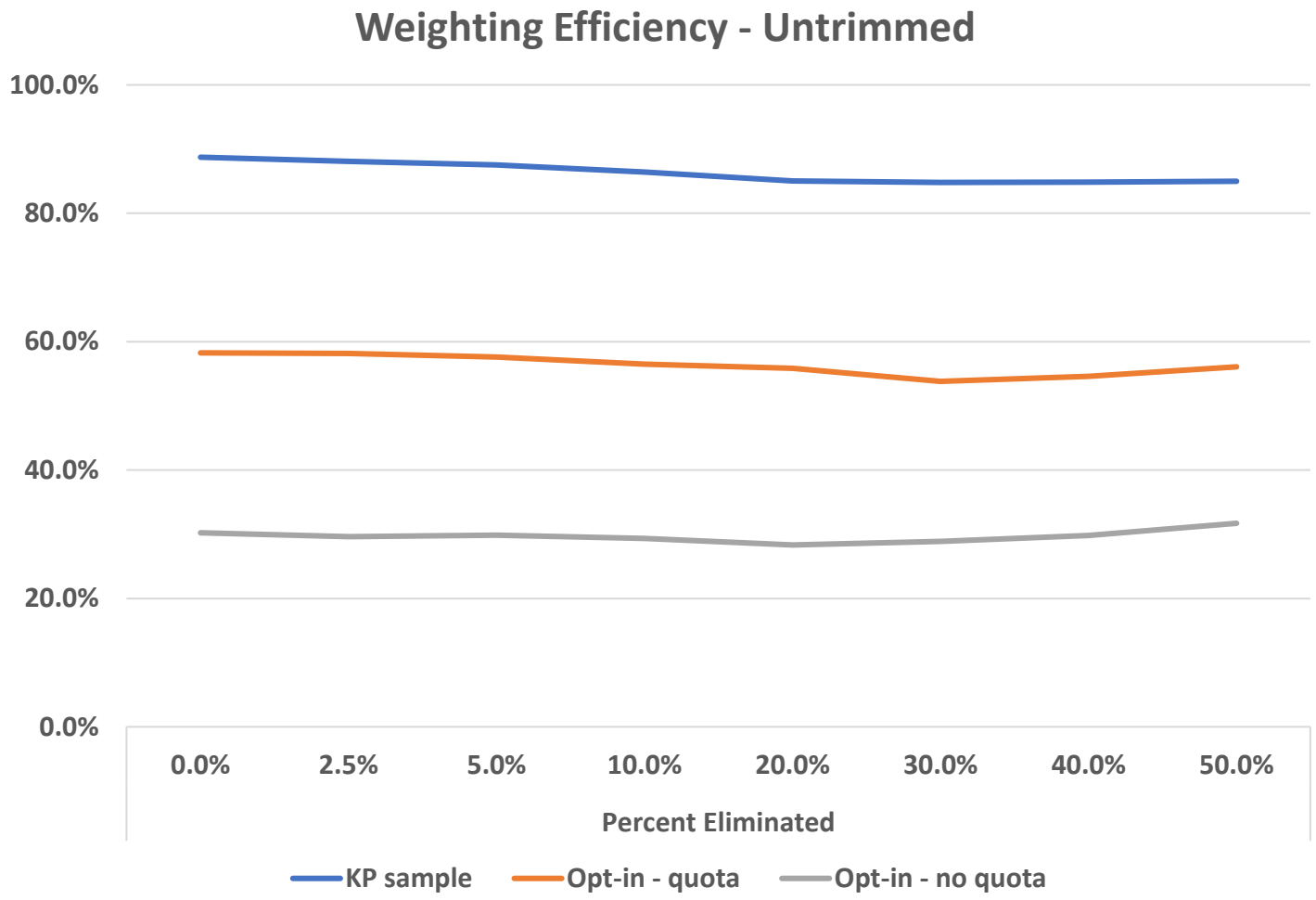
Comparing all sample sources, KnowledgePanel, bias was lowest for each race-ethnicity group. Across samples we found that the White and Black groups showed the lowest bias, while the Latino group had the highest bias. More cleaning did not generally reduce or increase bias for the race-ethnicity, though there may be some slight increase of bias with more extreme cleaning.

Study 2 Results – Effects of Cleaning on Weighting Efficiency - Untrimmed

Next, examined sample cleaning on weighting efficiency (which affects effective sample size) for each sample source using untrimmed weights (no constraints on low or high weights).

The KnowledgePanel probability-based panel had the highest weighting efficiency while the Opt-in No Quotas group had the lowest efficiency.

Cleaning did NOT improve weighting efficiency.

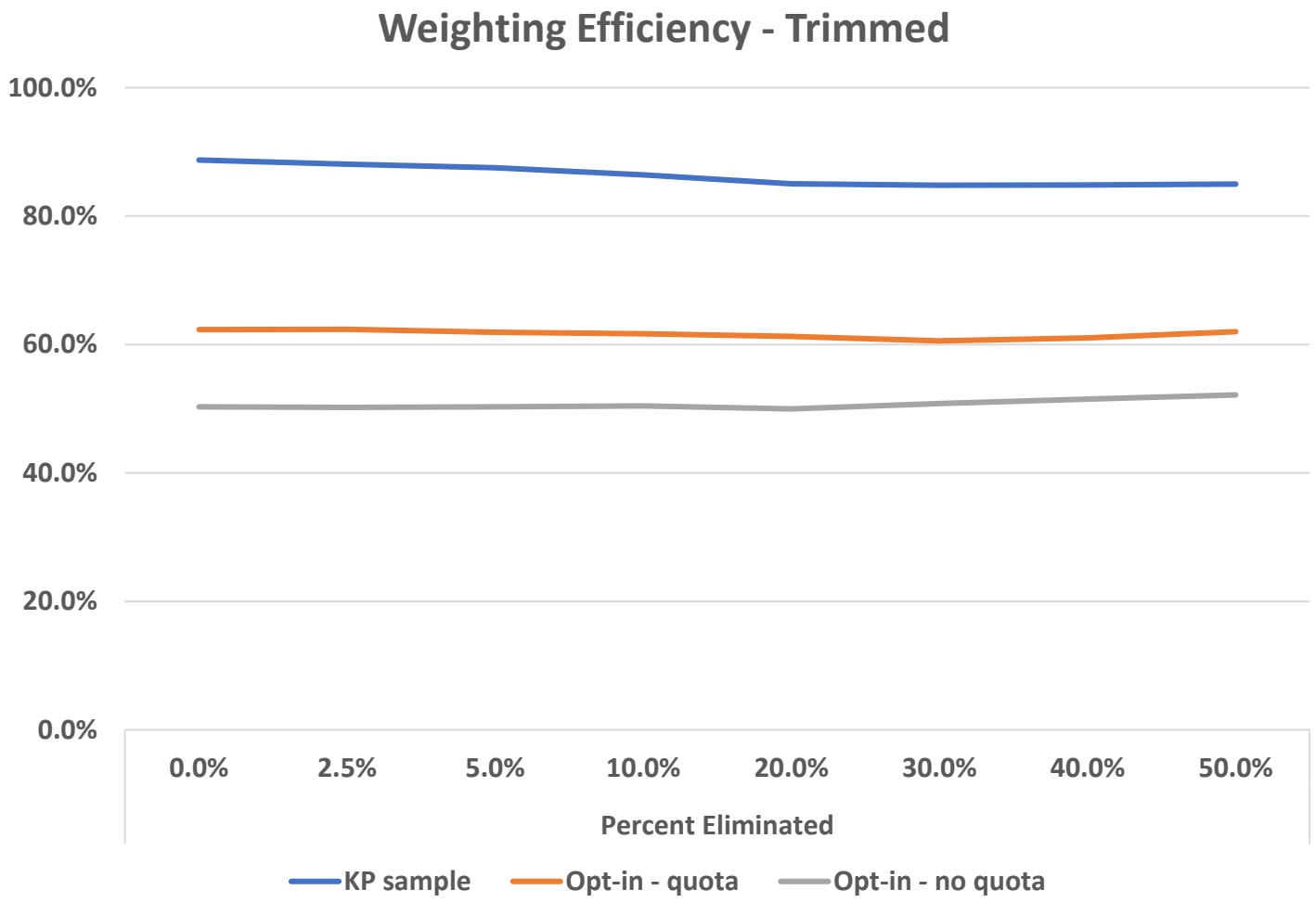


Study 2 Results – Effects of Cleaning on Weighting Efficiency - Trimmed

We then examined the weighting efficiency for trimmed weights, trimming weights to a range from .2 and 5.

KnowledgePanel weights did not need trimming (all weights were between .2 and 5) but trimming improved the efficiency of both opt-in samples.

Cleaning did NOT improve weighting efficiency for trimmed weights.



Discussion

Conclusions

- We anticipated a slight decrease in bias with minimal data cleaning and an increase in bias when removing larger proportions of respondents. However, we did not find any decrease or increase in bias with more rigorous cleaning.
- Results were confirmed in two studies with both KnowledgePanel and non-probability samples, using both a multiple cleaning criteria approach as well as just using speeding as the criterion to remove cases.
- Data cleanliness is NOT necessarily next to godliness – overall, and within specific race-ethnicity groups:
    - Data cleaning did not reduce bias
    - Cleaning did not change results
    - Cleaning does not appear to affect estimates for smaller groups – race-ethnicity.

Discussion

Why is the average bias/error unchanging no matter how many you eliminated?

- First, the fastest 1 or 2% or most egregious sub-optimal respondents (often, but not always the one and the same) do provide somewhat different responses than the other respondents. But eliminating the fastest 1 or 2% doesn’t change any overall point estimates (especially if the fastest are generally random responses or are similar to the other respondents' responses).
- Beyond the fastest 2% or most sub-optimal respondents, respondents who are faster do not significantly differ from slower respondents. Therefore, eliminating respondents based on speed beyond the fastest 1 to 2% eliminates people just like people who take longer to respond, again leading to little or real change in average estimates or bias, even if one eliminates the fastest 50% of respondents (though you do lose statistical power due to loss of respondents and increases in weight variance).

Next Steps

- **We are conducting additional research to assess how increased data cleaning could impact on variance and covariance of our benchmark measures – focus to date has been on point estimates (means and proportions).**
- **We will look to replicate these findings with additional groups (e.g., gender or age).**

Thank you!

Randall K. Thomas

Randall.Thomas@Ipsos.com