Federal Committee on Statistical Methodology

Scrubbed Clean: Does Data Cleaning Improve the Quality of Analytic Models?

Megan A. Hendrich, Randall K. Thomas, & Frances M. Barlas

Ipsos Public Affairs

2022 FCSM Research & Policy Conference
October 25th – October 27th, 2022
Washington D.C.

Study Background

Many researchers believe that it is necessary to clean survey data before analysis in order to improve data quality and accuracy. Sub-optimal response is believed to be a source of lower quality data due to dishonest, mistaken, inattentive, or approximate responses.

Data cleaning is often based on many sub-optimal behaviors:

- **Speeding through the survey**
- **Grid non-differentiation or straight-lining**
- **Item nonresponse (i.e., skipping items)**
- **Extreme responding on numeric entry**
- **Failure at trap questions (e.g., compliance traps)**
- **Consistency checks**

Study Background

While there is little research on data cleaning and its effects on measurement bias, what does exist seems to start with the assumption that data cleaning is necessary to improve the accuracy of survey results.

However, there are potential disadvantages to data cleaning:

- Takes time
- Has implications for sample costs
- May clean out harder-to-reach respondents more often

Study Background

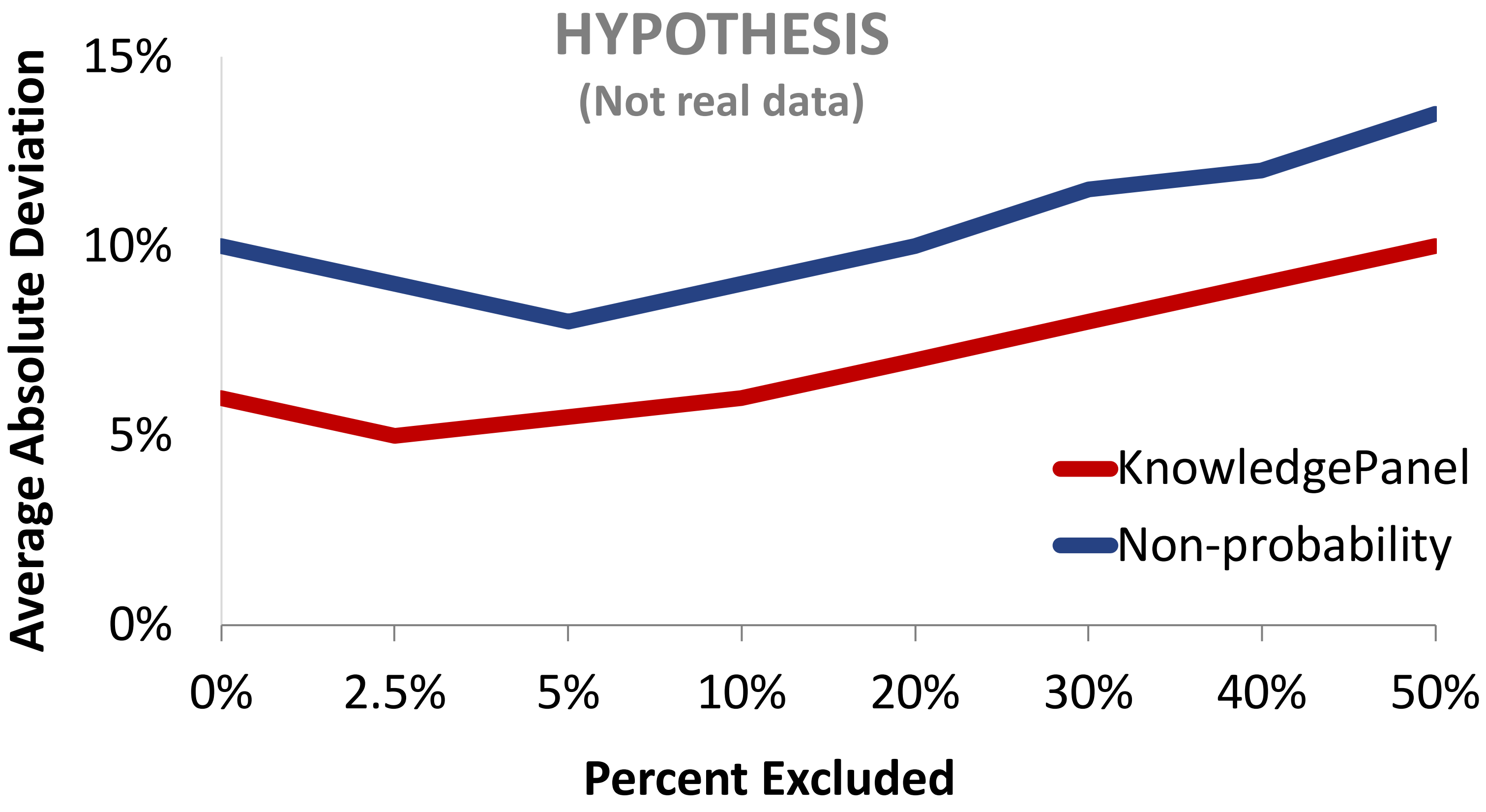
In an initial study using data collected from both probability and non-probability online samples, we used extensive cleaning criteria based on the following:

- **Item nonresponse**
- **Completion speed**
- **Grid non-differentiation**
- **Extreme numeric entry**

Using this cleaning criteria, we deleted cases in gradations from 2.5% up to 50%.

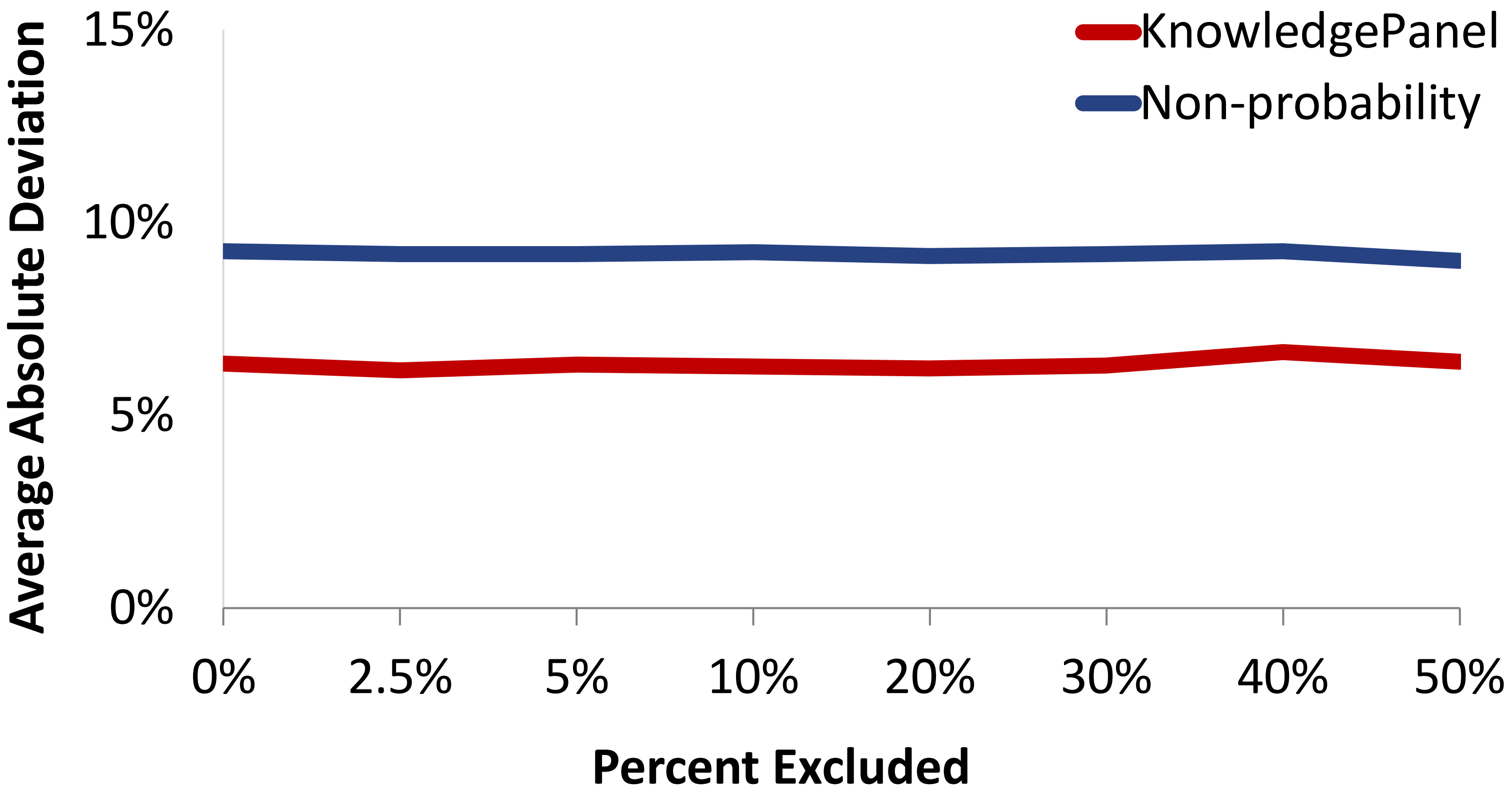
Study Background

We hypothesized that minimal data cleaning, around 2.5% to 5%, would reduce bias, but extensive cleaning would do more harm than good:



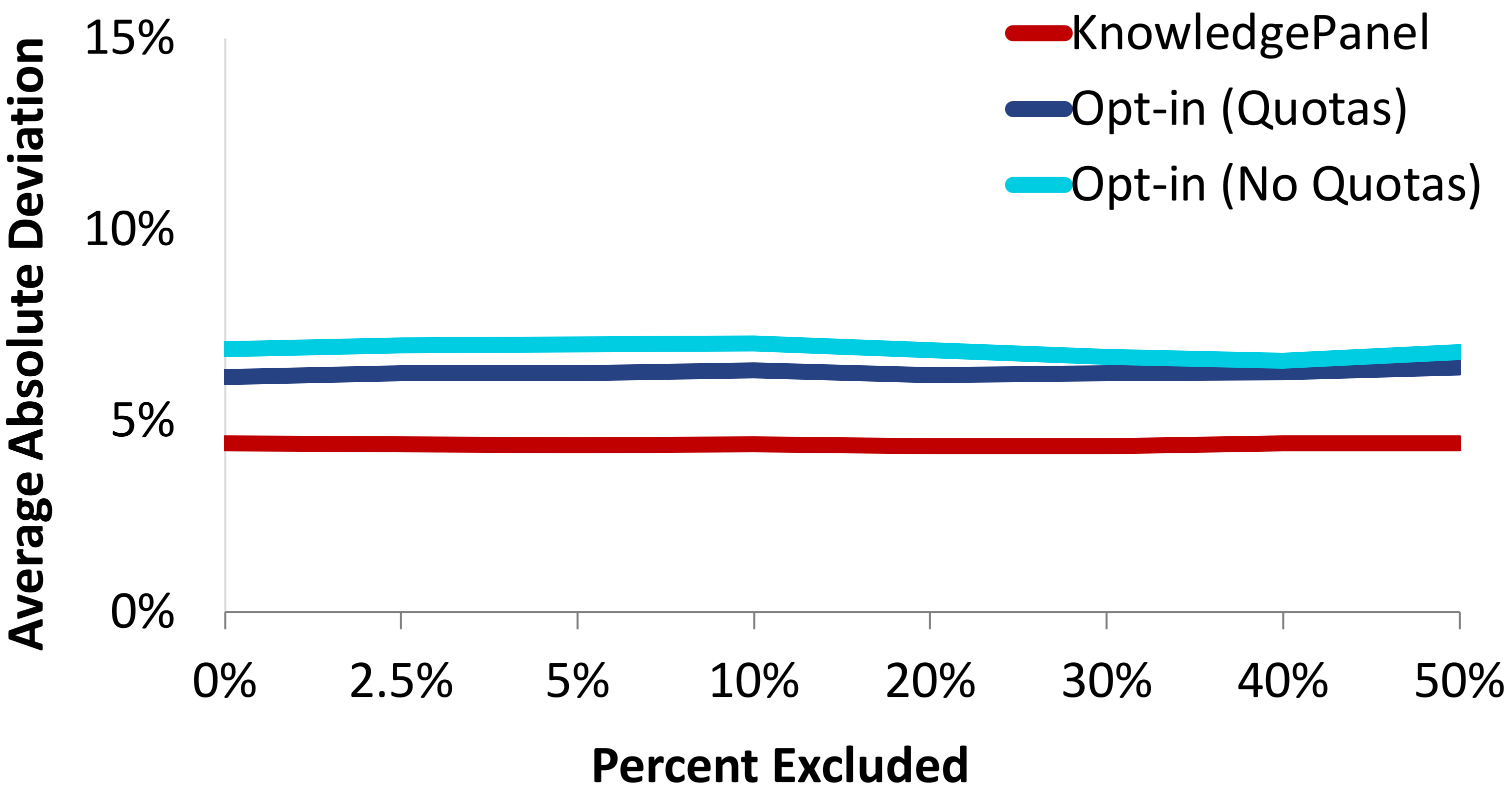
Study Background

However, we instead found that there was no effect on bias for point estimates with increasingly rigorous exclusion criteria:



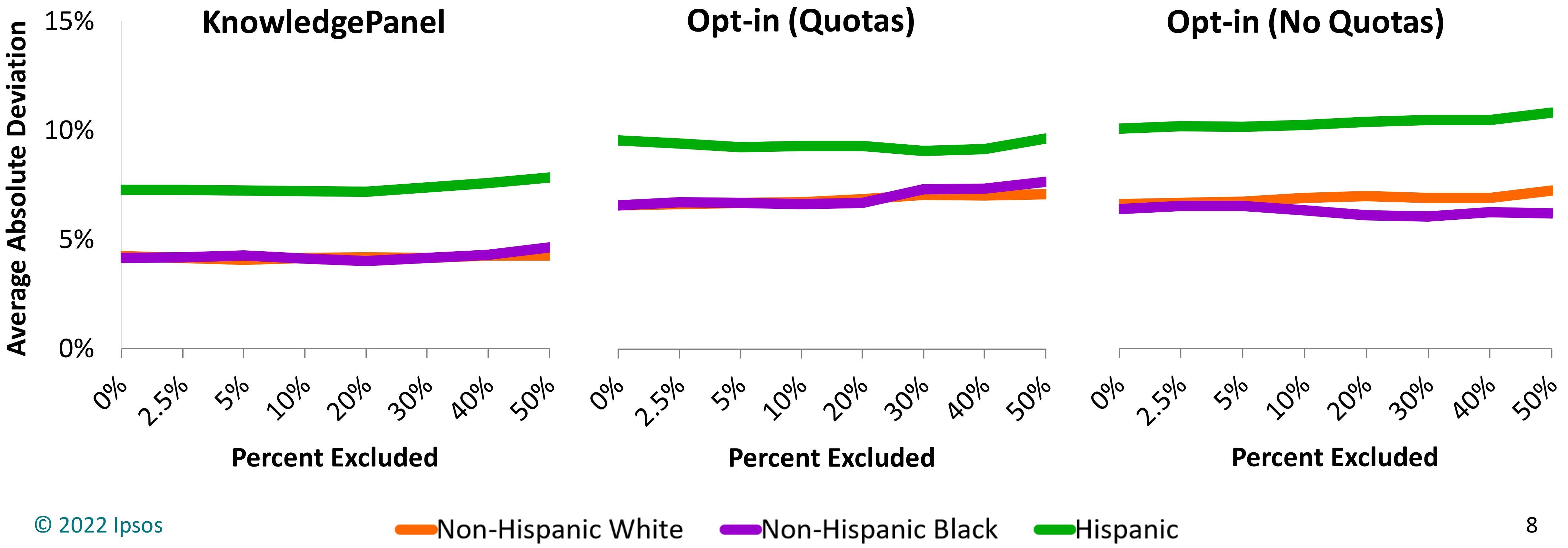
Study Background

In a follow-up study using only completion speed for cleaning, we again found no effect on bias with increasingly rigorous exclusion criteria:



Study Background

We also found that more data cleaning did not reduce or increase bias within race/ethnicity subgroups, though there may be a slight increase in bias with more extreme cleaning:



Study Purpose

Although no improvement has been found for point estimates (i.e., proportions, means) as a result of data cleaning, we were interested in finding out if data cleaning could affect covariance—specifically, correlational analyses using multiple regression.

In the current study, we sought to examine how regression models could be affected by varying degrees of data cleaning.

Method

Study Design

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

- **Ipsos KnowledgePanel: N = 3,344**
  - **The most well-documented, probability-based, online panel in the U.S. recruited primarily through address-based sampling**
- **Two non-probability samples:**
  - **Opt-in using quotas for gender by age, race/ethnicity, and education to obtain a demographically balanced sample: N = 2,677**
  - **Opt-in without quotas: N = 3,293**

Study Design

Data cleaning method:

- **We used completion speed as the primary criterion for cleaning.**
- **We created groups within each sample type that eliminated 0%, then the fastest 2.5%, 5%, 10%, 20%, 30%, 40%, and 50% of each sample.**

We then ran two regression models for each dataset under the varying levels of data exclusion due to speeding.

Analytic Design – Model 1

Model 1 – Dependent variable: Life satisfaction (1=Not satisfied; 5=Completely satisfied); Predictors:

- **Positive emotions**
- **Negative emotions**
- **Quality of healthcare**
- **Quality of places to live**
- **Quality of education**
- **Quality of jobs available**
- **Self-rating of health**

Analytic Design – Model 2

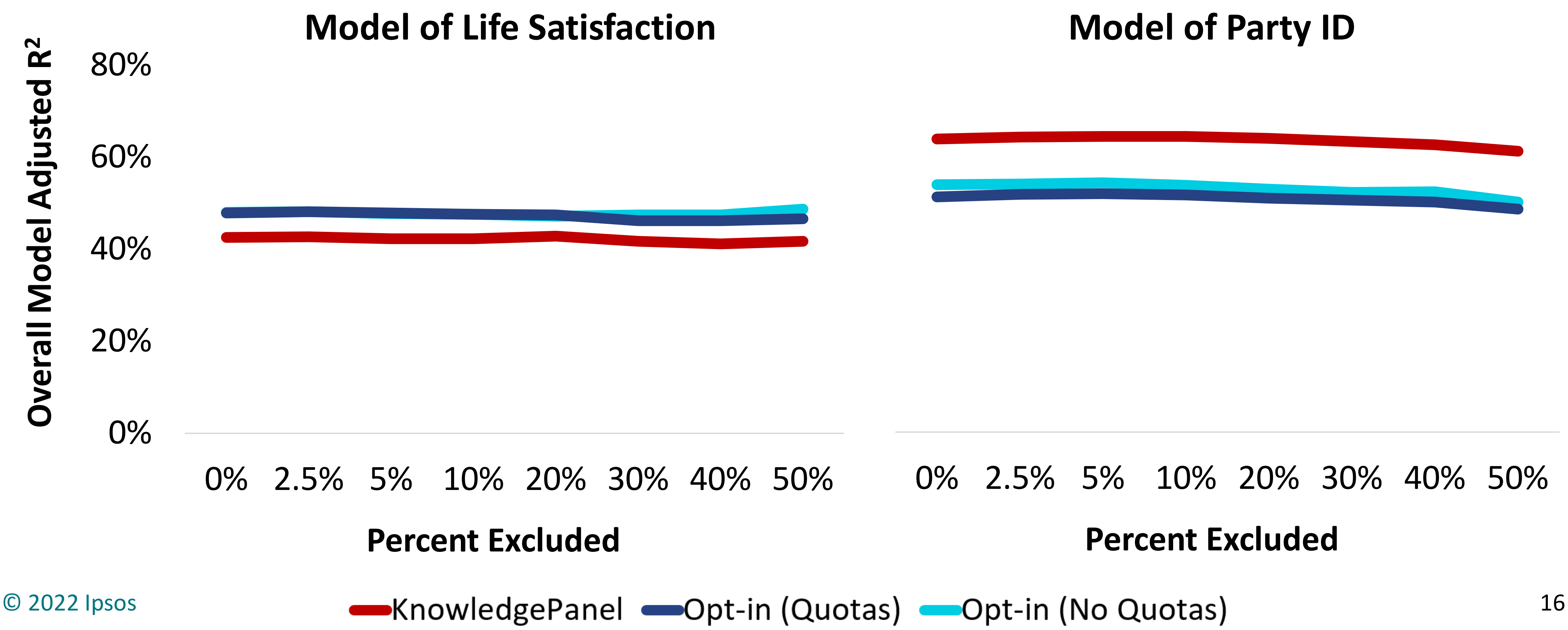
Model 2 – Dependent variable: Political party identification (1=Strong Republican; 7=Strong Democrat); Predictors:

- Protect gun ownership
- Government should do more for environment
- Government spending too much for Black persons
- Abortion should be illegal
- Government should provide healthcare for all
- Government should reduce the wealth gap
- Government should increase military spending
- Support for Black Lives Matter
- Allow illegal immigrants to be citizens

Results

Results – Overall Model Adjusted R^2

We did not find any differences or improvement in the amount of variance predicted (Adjusted R^2) as more cases were deleted.

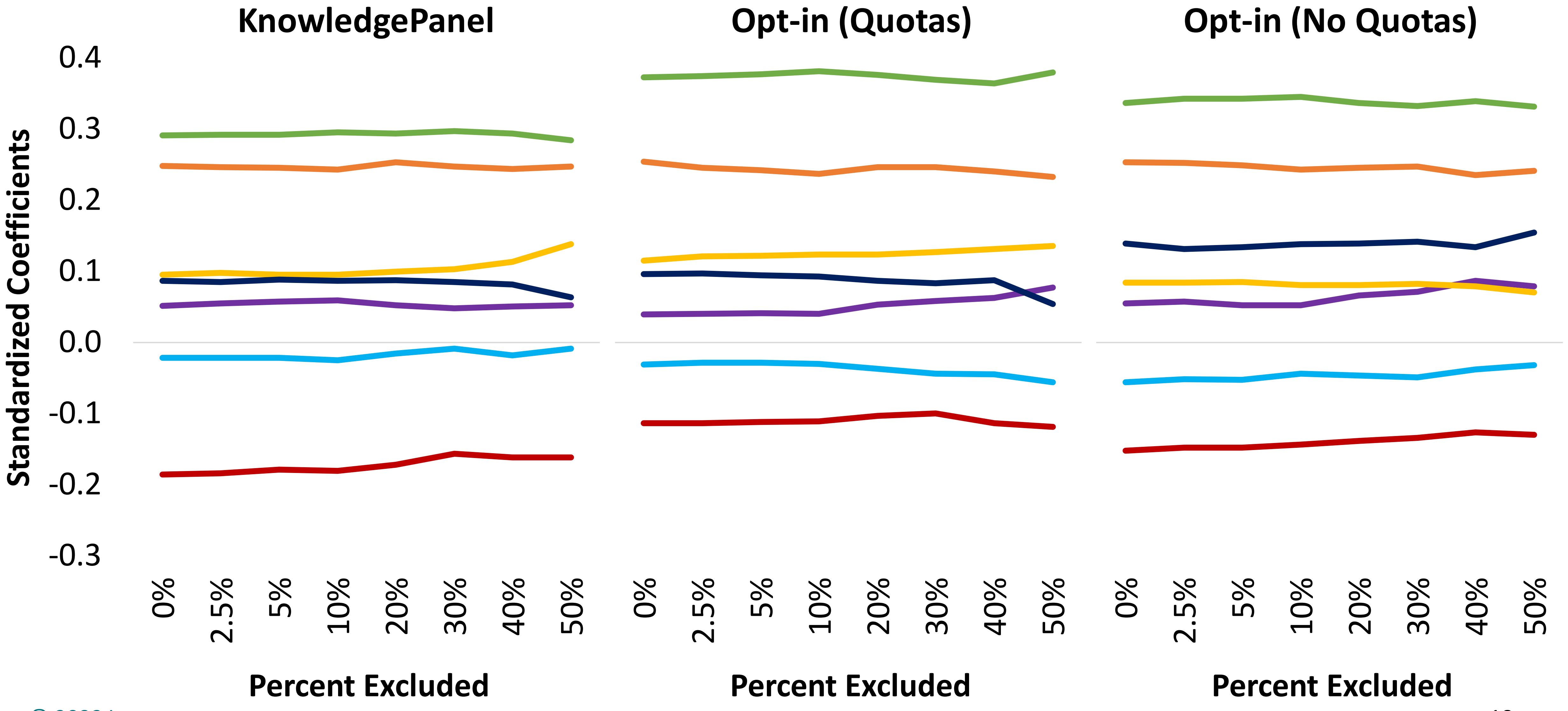


Results – Beta Coefficients

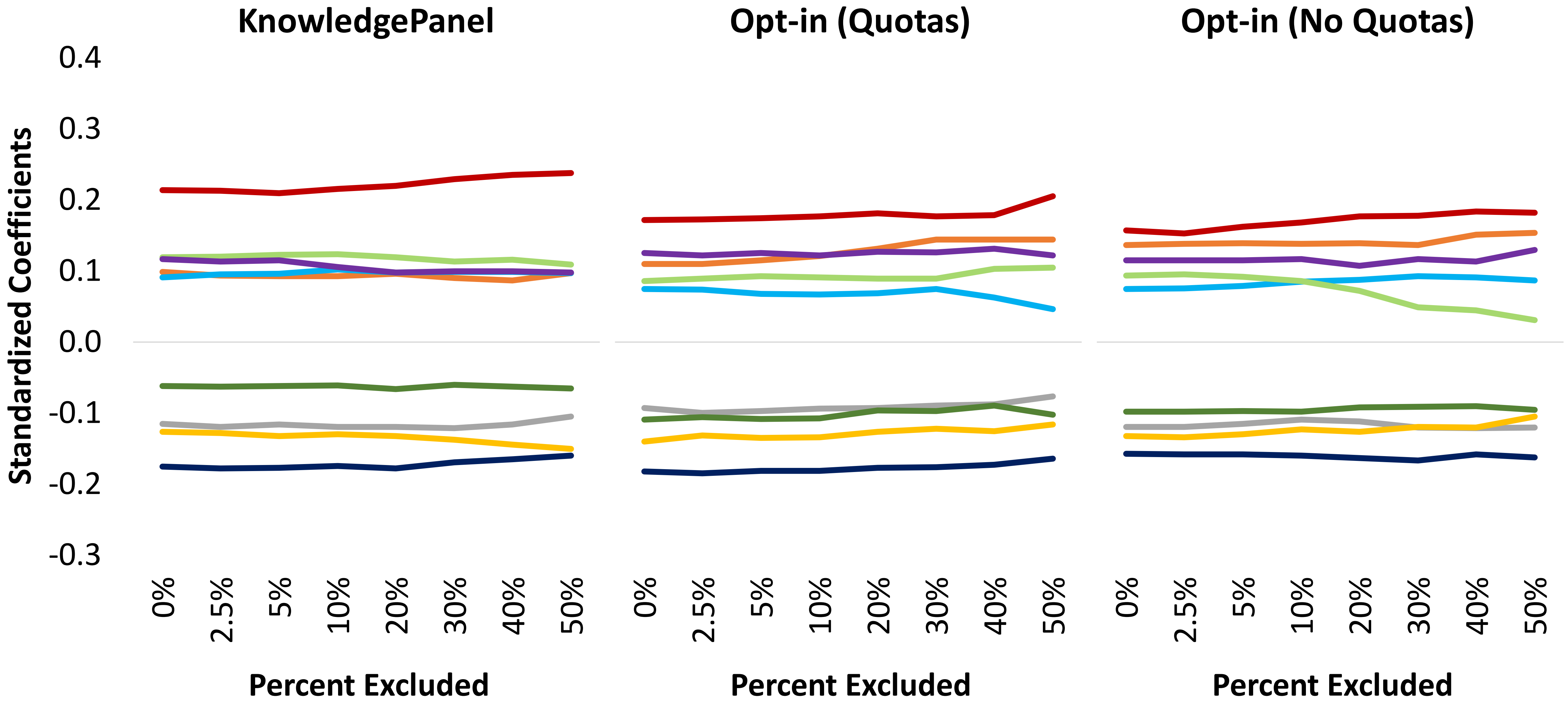
While the overall R^2 may not have changed much with increased data cleaning, it is possible that as more cases are deleted, the predictor coefficients in the form of the betas could become more unstable and show divergence with increased cleaning.

However, for both models, we did not see any systematic changes in betas as increased cleaning was performed, though some betas became more unstable with extreme levels of data cleaning (around 30% deletion or higher).

Results – Beta Coefficients (Model 1: Life Satisfaction)



Results – Beta Coefficients (Model 2: Party ID)



Discussion

Conclusions and Discussion

If data cleaning eliminated 'noise' in the data from sub-optimal response, we would expect that at least some cleaning would improve the correlations between variables, when correlations existed between the variables. However, we did not find any evidence to support this.

Similar to our findings regarding no reduction of bias from standard benchmarks, we did not find that data cleaning improves model validity in terms of improving the overall model predictive utility.

Conclusions and Discussion

Why are correlational models not improved, no matter how many respondents we eliminated?

- **The fastest 1-2%, or the most egregious sub-optimal respondents, do provide somewhat different responses than other respondents; however, eliminating them doesn't change the overall point estimates, variances, or covariances (especially if the fastest are generally random responses or are similar to other respondents' responses; Thomas, 2014).**

- **Beyond the fastest 2%, other faster respondents do not significantly differ from slower respondents. Therefore, cleaning out these faster respondents eliminates people who are just like those who take longer to respond, leading to little to no change in the estimates (though you do lose statistical power due to loss of respondents and increases in weight variance).**

Thank you!

Megan A. Hendrich

Megan.Hendrich@ipsos.com

Appendix: Coefficient Legends

Model 1: Life Satisfaction

- Positive emotions
- Negative emotions
- Quality of healthcare
- Quality of places to live
- Quality of education
- Quality of jobs available
- Self-rating of health

Model 2: Party ID

- Protect gun ownership
- Government should do more for environment
- Government spending too much for Black persons
- Abortion should be illegal
- Government should provide healthcare for all
- Government should reduce the wealth gap
- Government should increase military spending
- Support for Black Lives Matter
- Allow illegal immigrants to be citizens