Analyzing Survey Duration Time and Its Relationship to Data Quality for a Household Survey

Federal Committee on Statistical Methodology

October 27, 2022

By

S. Grace Deng, L. Kaili Diamond

Disclaimer: The analysis and conclusions contained in this presentation are those of the authors and do not represent the official position of the U.S. Energy Information Administration or the U.S. Department of Energy.

Introduction

- Residential Energy Consumption Survey (RECS)
  - Collects energy-related characteristics and energy consumption billing data from the Household Survey and the Energy Supplier Survey
  - 2020 RECS Household Survey was conducted entirely in self-administered web and mail modes: total of 18,496 responding households, 73% via web and 27% via paper

- Survey duration time
  - The length of time it takes a respondent to complete the set of questions in the Household Survey questionnaire
  - Both the total duration time in completing the entire survey and the duration time in completing each section were collected as paradata in the web instrument
  - 11,328 web cases in this analysis

Research questions

1. How is the survey duration time related to respondent demographics and housing characteristics?
   - e.g.: housing type, respondent age, respondent education

2. How is the survey duration time related to the data quality of survey responses?
   - e.g.: percentage of missing questions, batch-edit changes

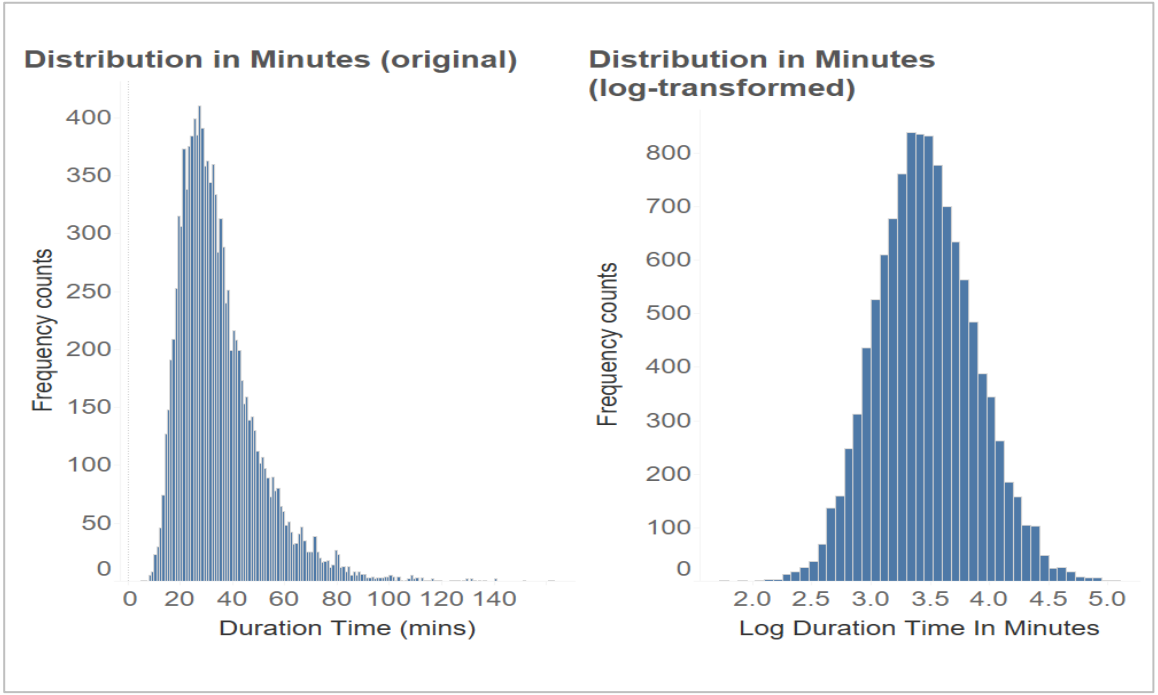
General linear model was used to check if demographic and housing characteristics are correlated with duration time

Predictors in the model:

- **Housing type (3)**: mobile, single family (attached and detached), multifamily (apartments at least 2 units)
- **Total household members (4)**: 1, 2, 3-4, >=5
- **Respondent education (4)**: High school or below; some college; bachelor; master or above
- **Respondent age (5)**: below 25; 25-34; 45-54; 55-64; 64+
- **Respondent sex (2)**: male; female
- **Employment (4)**: full-time; part-time; retired; not-employed

Dependent variable duration time was log transformed

Duration time was log-transformed due to its skewed distribution

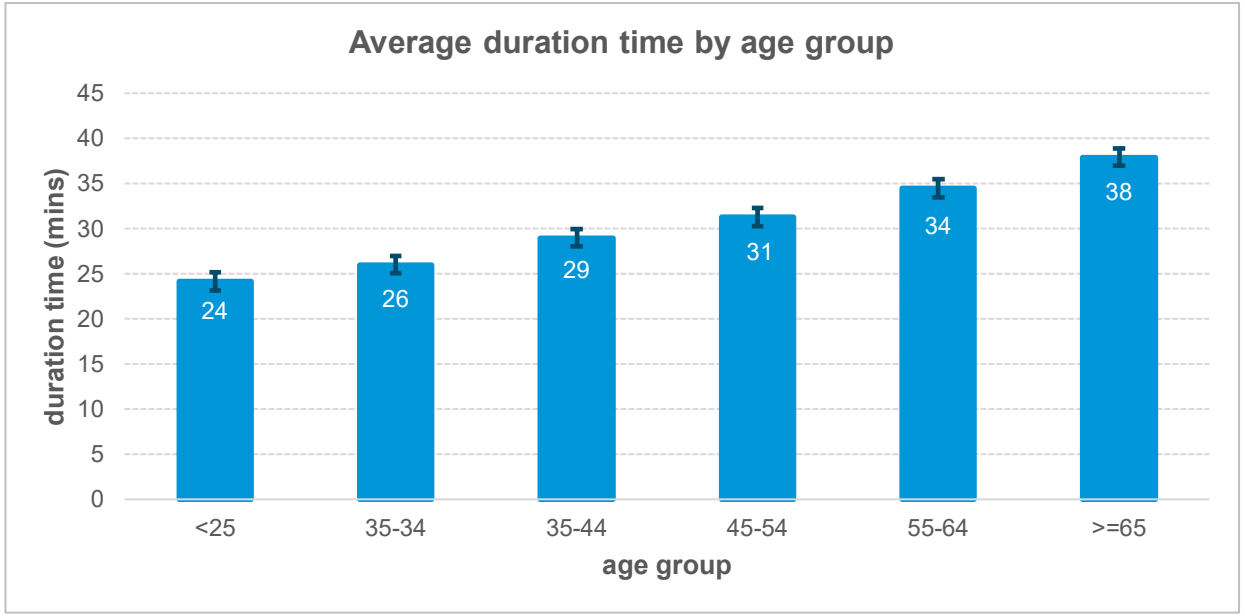


Respondent age had the largest effect on the average duration time, other effects were relatively small

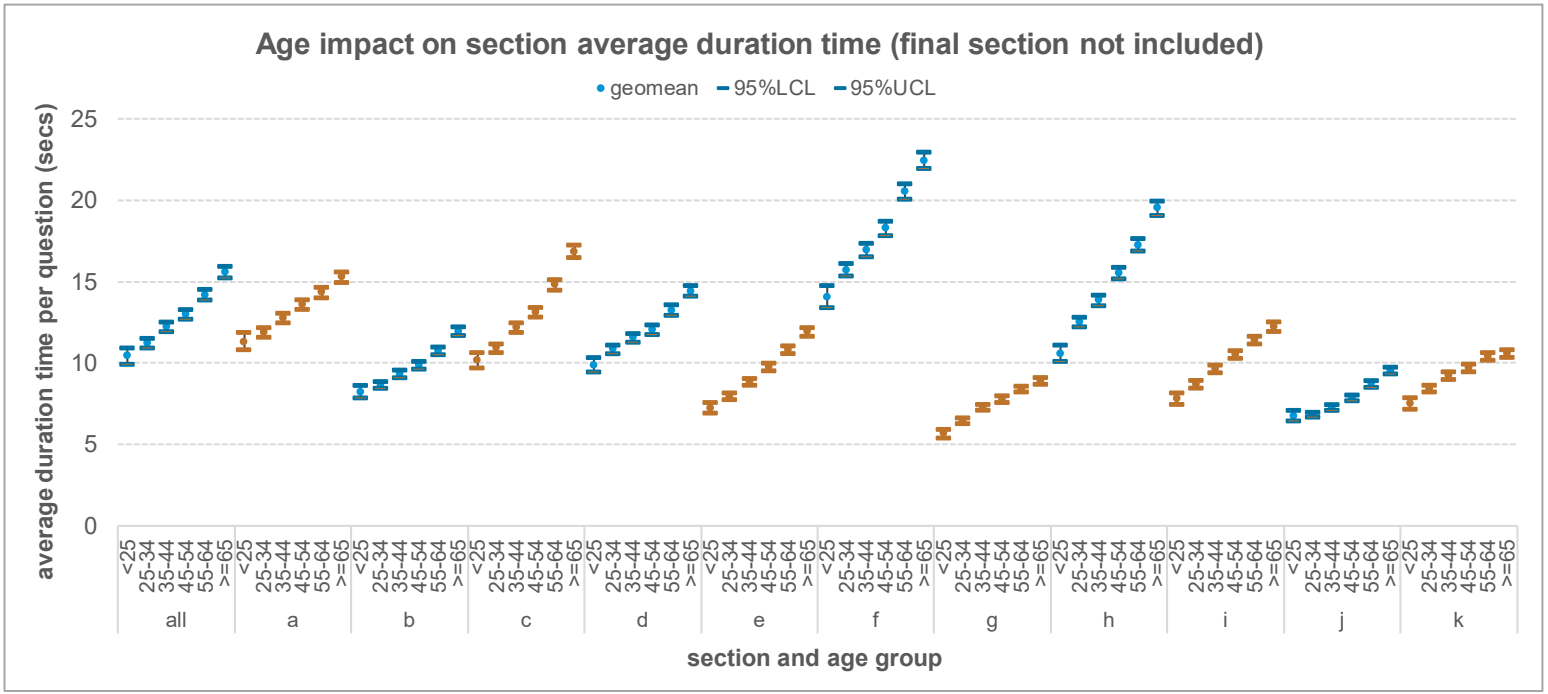
Source	Degree of freedom	Sum of Square	Mean Square	F Value	Pr > F
respondent age	5	122.8	24.4	167.6	<.0001
respondent employment	3	10.3	3.4	23.6	<.0001
total household members	3	9.1	3.0	20.9	<.0001
housing type	2	4.1	2.1	14.2	<.0001
respondent education	3	3.8	1.3	8.8	<.0001
respondent sex	1	3.3	3.3	22.6	<.0001
housing type*education	6	2.0	0.3	2.3	0.0294
education*sex	3	1.8	0.6	4.1	0.0065
total member*sex	3	1.5	0.5	3.5	0.0148

If separated by housing type, respondent age still have the largest effect.

Overall, the older the age, the longer the duration time it took respondents to complete the survey

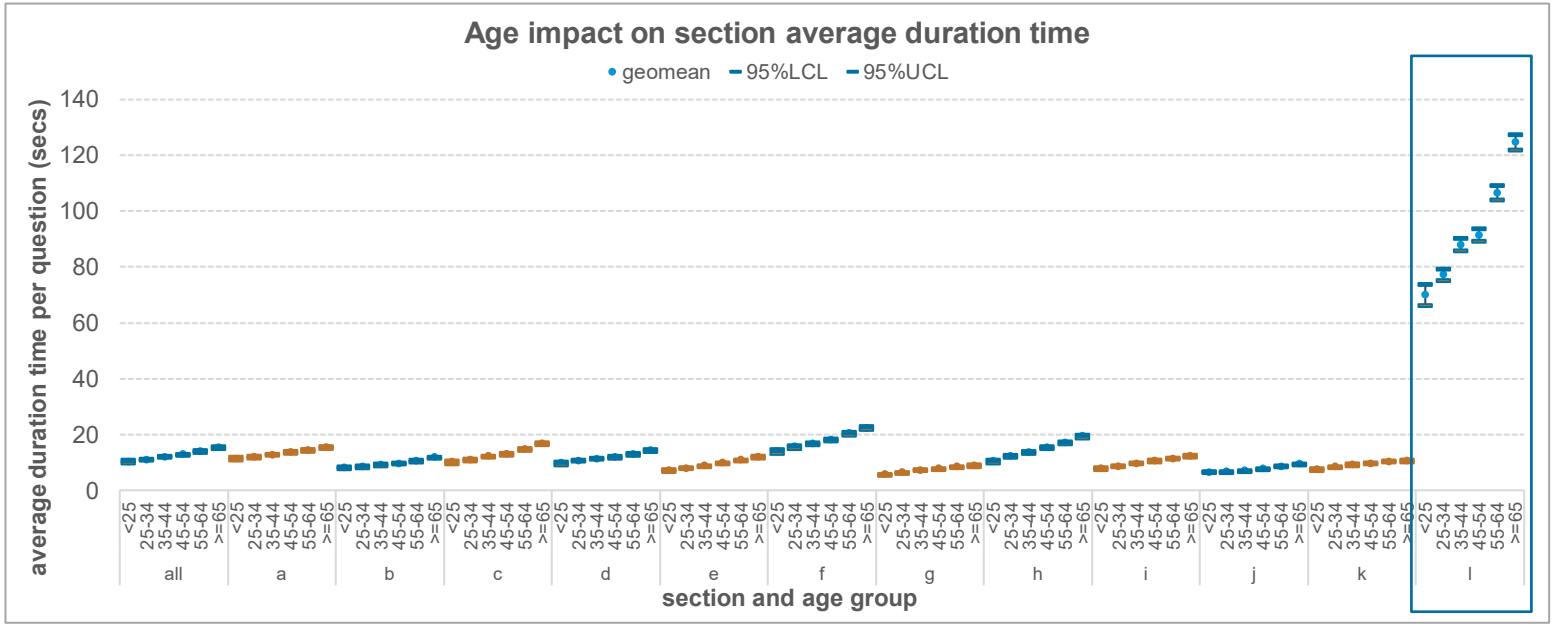


The age impact had the same systematic trends across all sections



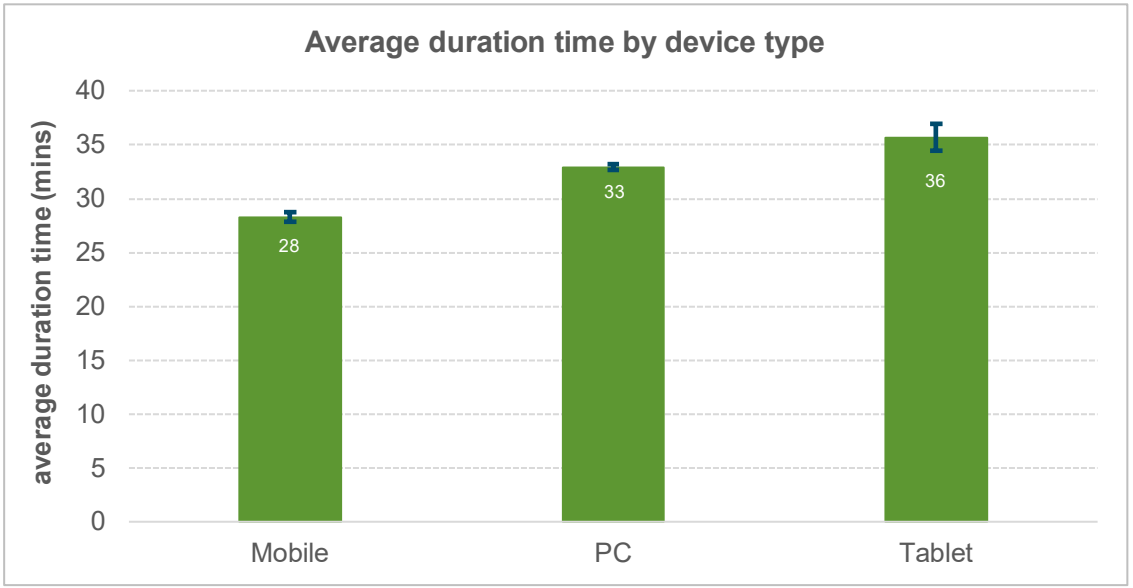
* Color scheme here is only for section readability

The average time on the supplier information section took significantly longer than the rest of the sections



* Color scheme here is only for section readability

Mobile phone users took the shortest average time in completing the survey, and device type was highly correlated with age



Evaluating data quality by three "speed groups"

Defining baseline group, fast group, and slow group for duration time

- **Fast group**: the lower 5th percentile from each housing type category
- **Baseline group**: between the 2nd and 3rd quantiles from each housing type category
- **Slow group**: the upper 5th percentile from housing type category

Metrics for data quality evaluation

- Percentage of missing questions
- Percentage of missing energy supplier information
- Percentage of variables with explicit don't know responses
- Batch-edit changes (automatic corrections to variables)
- Analyst-edit changes (manual corrections to variables)

Statistical analysis

- Generalized linear models/logistic regression models

Missing questions

A RECS survey question can include multiple response variables, and the number of eligible questions vary among respondents due to skip patterns.

Missing questions: all of the response variables in a question are missing

- Skipped questions were not counted

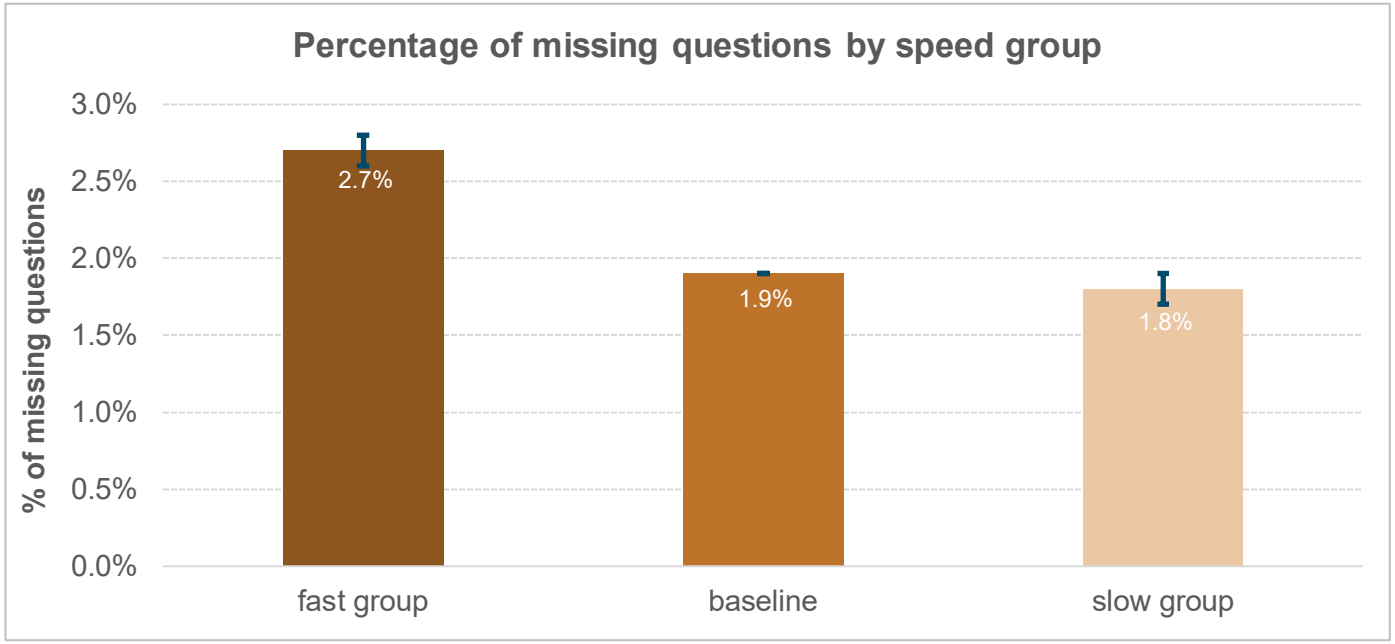
Example: one question with six response variables

In your home, which of the following types of computer equipment are used for teleworking or working from home?

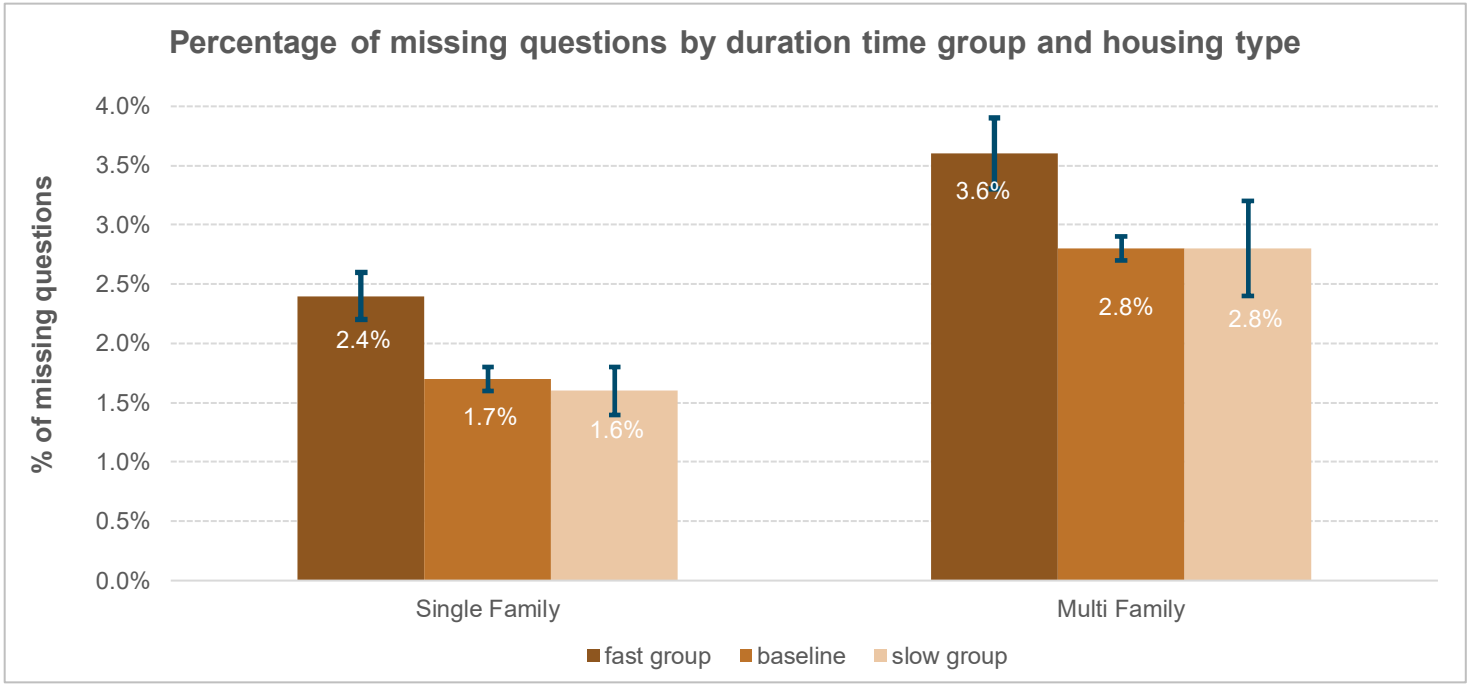
Desktop computer? **[TLDESKTOP]** (Yes/No)
Laptop computer? **[TLLAPTOP]** (Yes/No)
Tablet? **[TLTABLET]** (Yes/No)
External monitor? **[TLMONITOR]** (Yes/No)
Other **[TLOTHER, TLOTHER_other]** (Yes/No)

Section	# Total survey questions	Total survey variables
a - Your Home	39	62
b - Appliance	45	55
c - Electronics	26	59
d - Space heating	21	31
e - Air Conditioning	15	21
f - Thermostats and Temperatures	5	11
g - Water Heating	8	11
h - Lighting	7	11
i - Energy Bills	25	49
j - Household Characteristics	10	18
k - Energy Assistance	14	25
l - Final Questions	4	20
	219	**373**

Overall, the fast group had significantly higher average percentage of missing questions than those of the baseline and slow groups (~2%)



Within housing type, the fast groups in both SF and MF also had higher percentage of missing questions

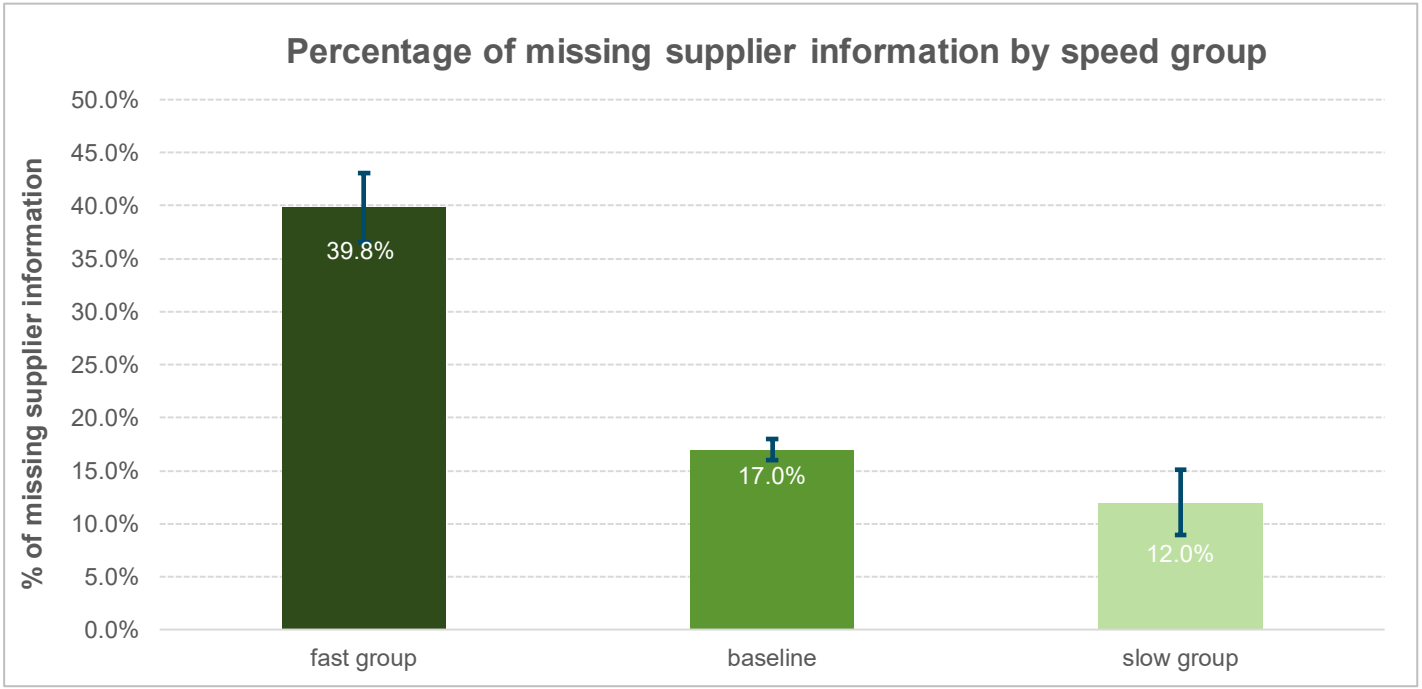


Missing energy supplier information

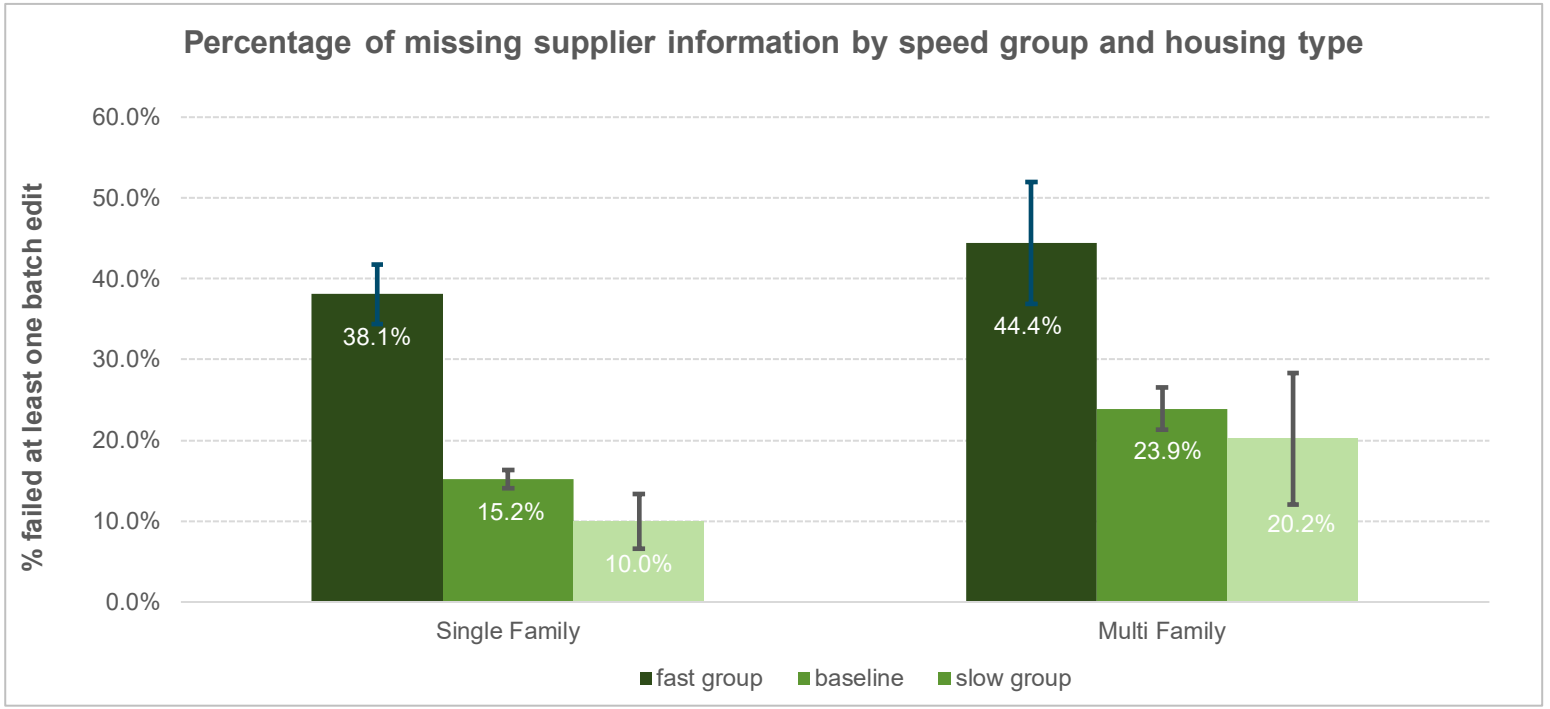
Survey asks for supplier information on electricity, natural gas, fuel oil, and propane in separate questions, if a respondent provided at least one fuel supplier, then it is considered a response.

Missing energy supplier information: no supplier information was provided for any of the fuel questions

The fast group had higher percentage of missing energy supplier information



Within housing type, the fast groups in both SF and MF had higher percentage of missing energy supplier information



Answered explicit "don't know"

Some questions offered explicit "don't know" responses. There is a total of 49 response variables (skips were not counted)

Example:

About how many windows does your home have?

1 1 or 2 windows

2 3 to 5 windows

3 6 to 9 windows

4 10 to 15 windows

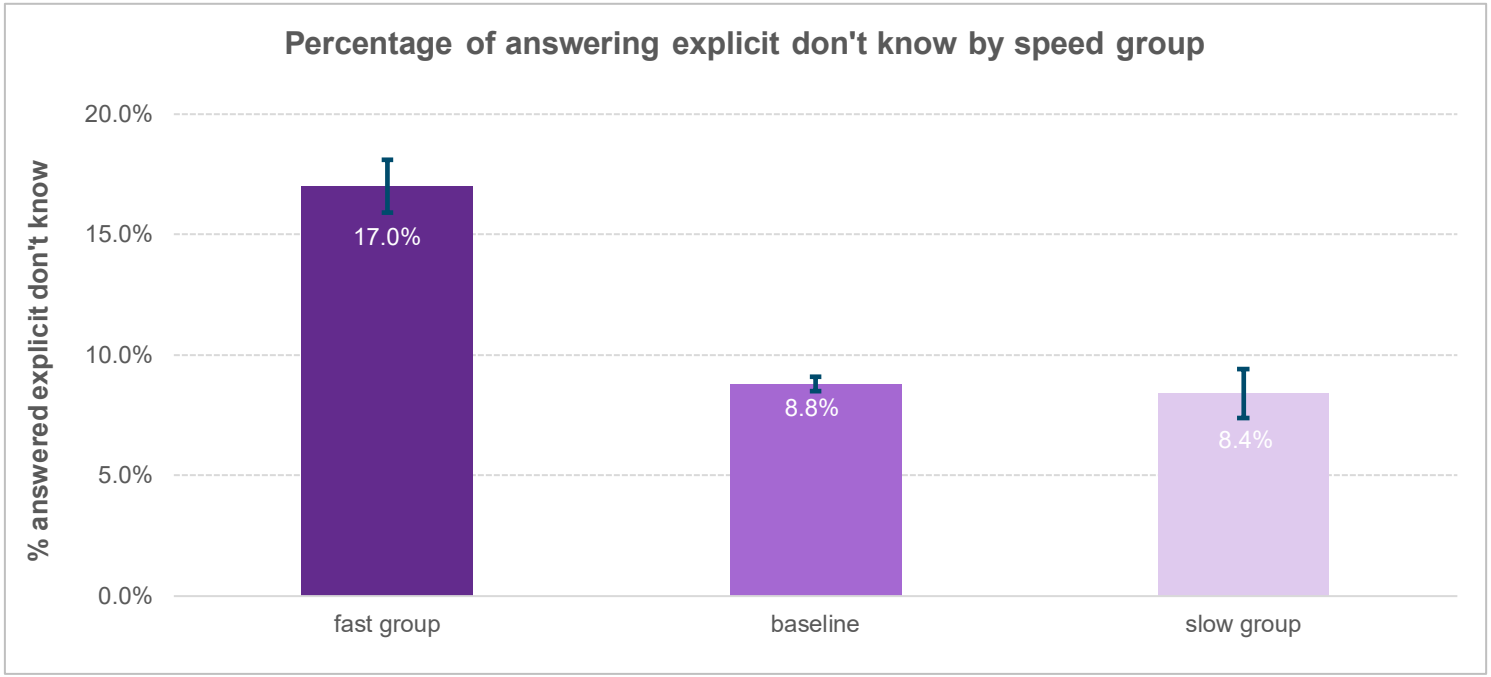
5 16 to 19 windows

6 20 to 29 windows

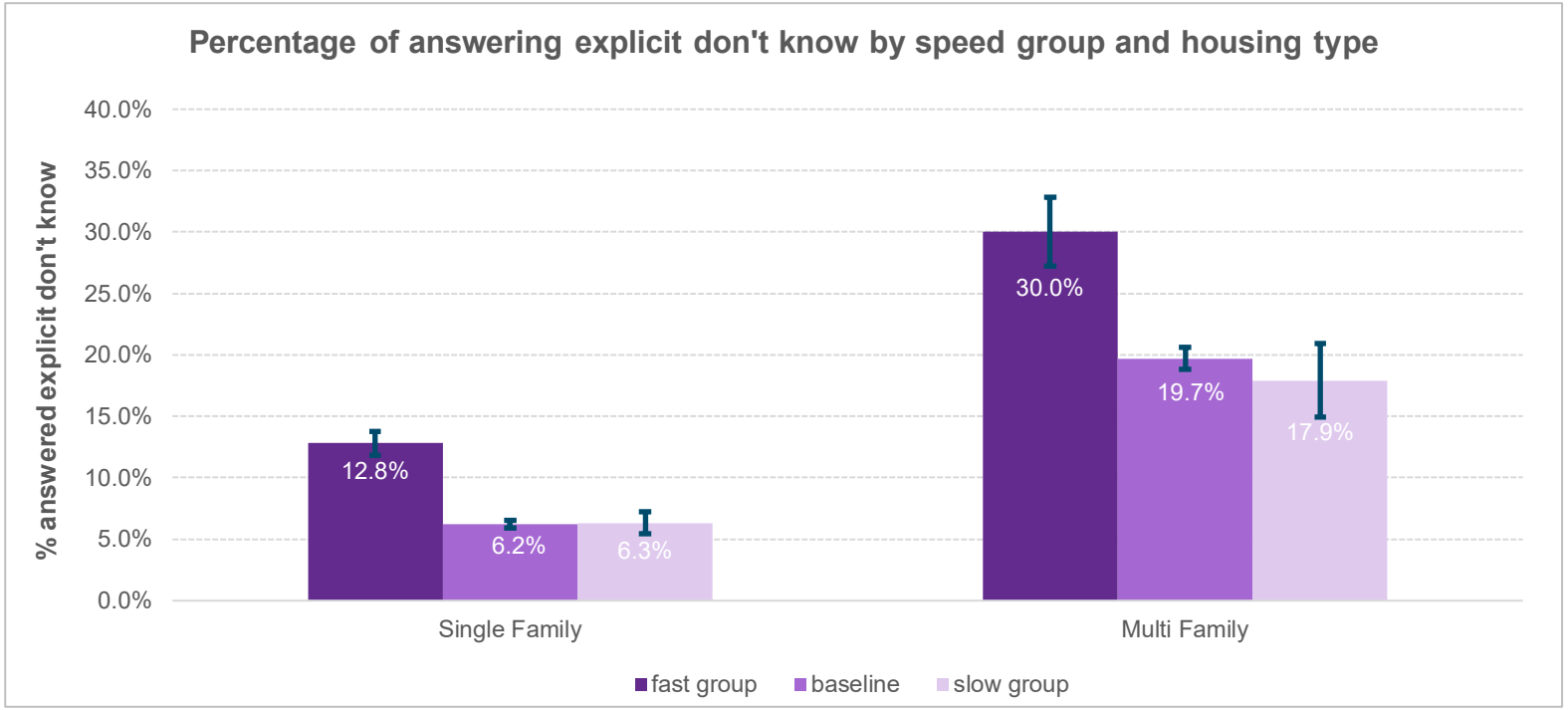
7 30 or more windows

-4 Don't know

The fast group had higher percentage of answering explicit don't know



Within housing type, the fast groups in both SF and MF had higher percentage of answering explicit don't know



Batch-edit changes

These are automatic corrections to response variables, includes:

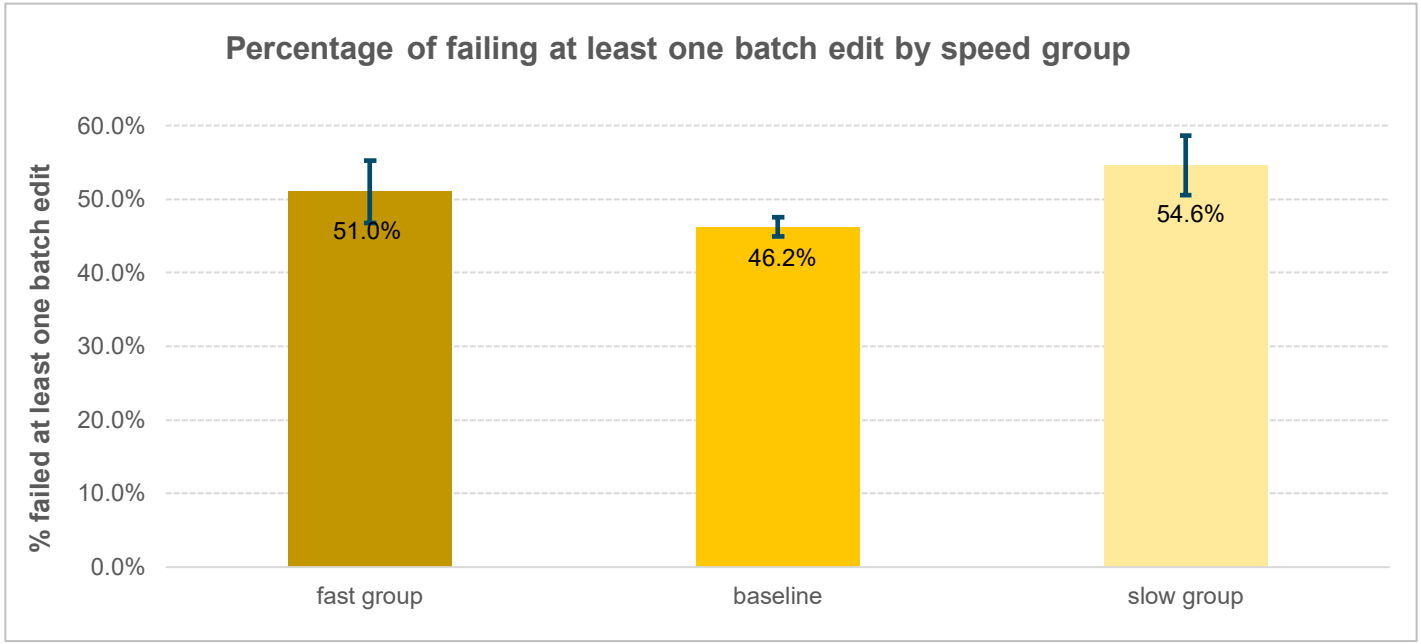
- Setting illogical outlier values to missing (e.g., TV is used more than 24 hours a day)

- Setting inconsistent responses to missing (e.g., reporting heating equipment as portal electric heater but saying the fuel is natural gas)

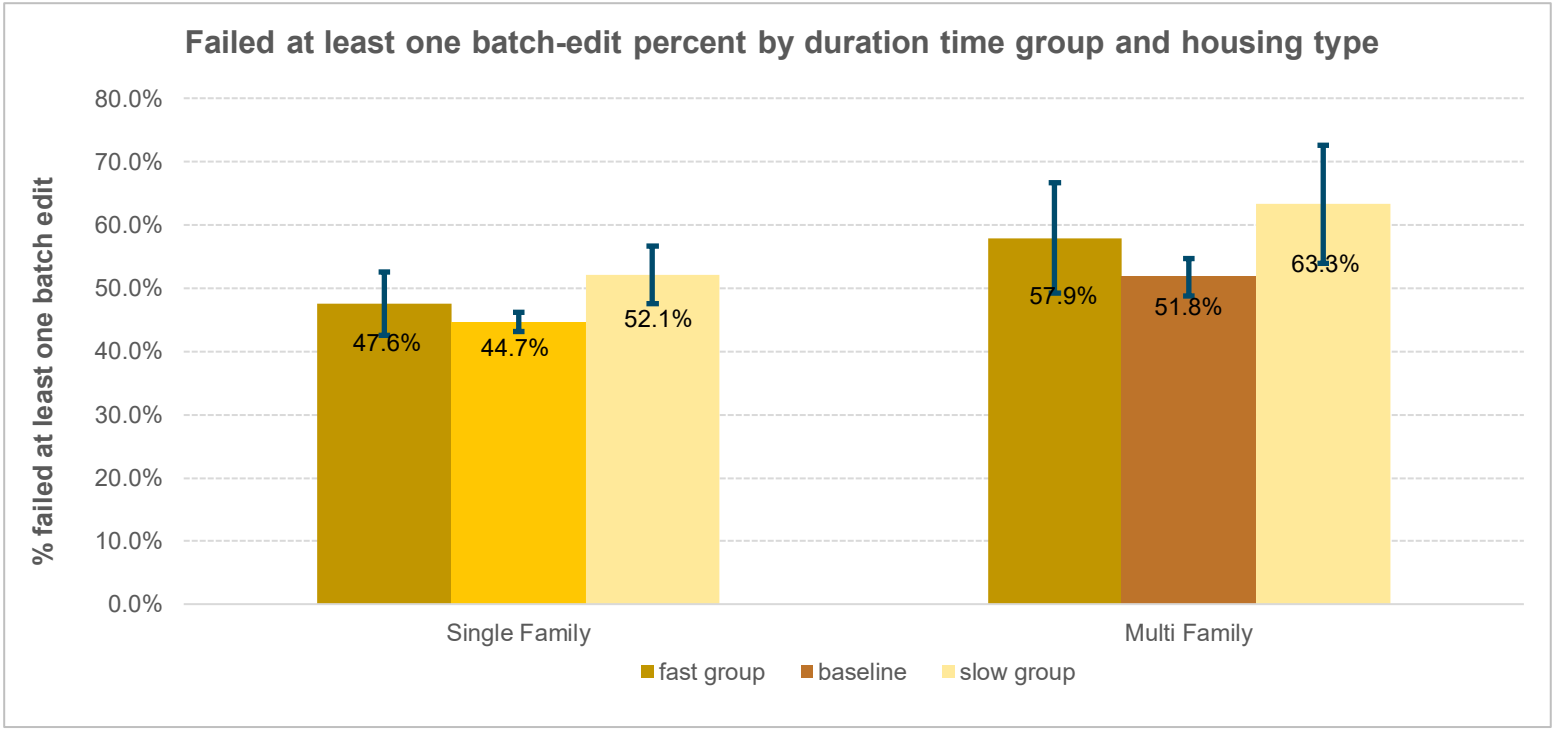
Analyst-edit changes

These are cases required analyst reviews and making manual corrections

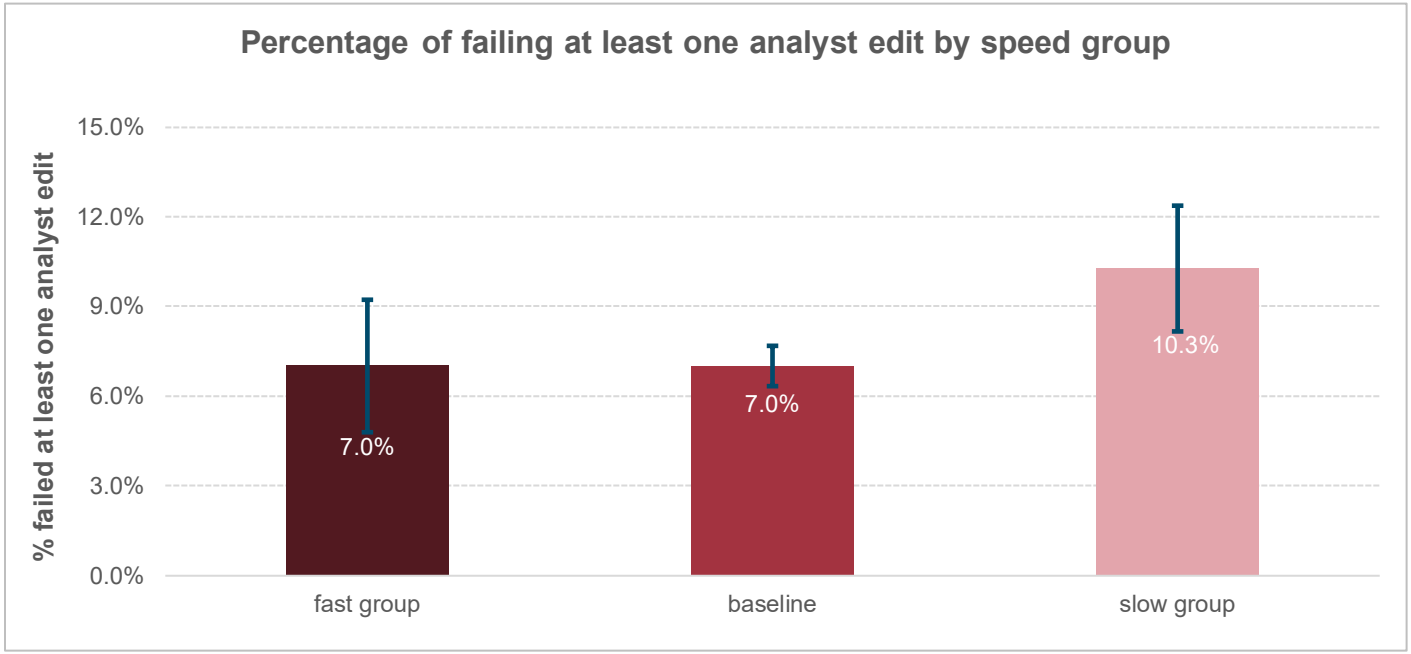
For batch-edit changes, the fast group was not statistically different from the baseline group, but the slow group was higher compared to that of baseline



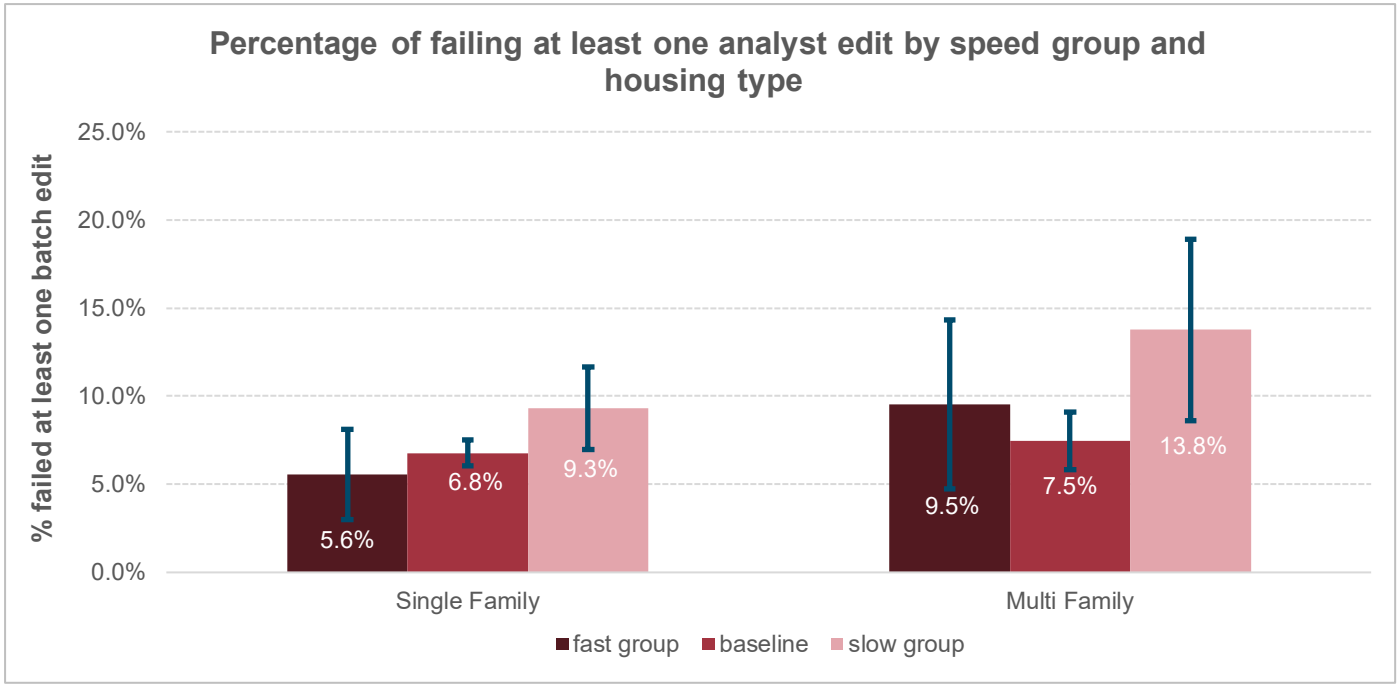
Within housing type, the trend in batch-edit changes were the same as the overall



Similarly for analyst-edit changes, the fast group was not statistically different from the baseline group; but the slow group was higher



Within housing type, no statistical differences were found in analyst-edit changes among speed groups



Conclusion:

- Among different characteristics, respondent age had the largest impact on duration time. The other characteristics effects were relatively small. Device type was highly correlated with age, and mobile phone users took shorter time than those of the PC or Tablet users.

- Duration time could also be affected by the nature of the survey questions (e.g., reporting supplier information would require longer time).

- The fast group that took shorter time were more likely to leave the questions blank and choose explicit don't know than the baseline or slow groups. However, in terms of batch-edit or analyst-edit changes, comparing to the baseline, this group did not have more edit changes, but the slow group had significant higher edit changes. These would be due to the fast group are more likely not to respond when they don't know the answer.

Future work:

- Additional editing from reconciliation between household characteristics and energy bills are still ongoing, we are planning to include these edits for future analysis.

Thank You!

- Grace Deng|Shaofen.deng@eia.gov
- Kaili Diamond|Kaili.diamond@eia.gov

BACKUP SLIDES

Speed group time definition and sample size

Speed group	Mobile	SF	MF
Fast group	<=16mins	<=17mins	<=15mins
Baseline	≥23 & ≤40mins	≥25 & ≤43mins	≥21 & ≤38mins
Slow group	≥64mins	≥65mins	≥64mins

Speed group	Mobile	SF	MF	Total
Fast group	23	378	126	**527**
Baseline	238	4372	1099	**5709**
Slow group	24	451	109	**584**