

# Estimating medians for combined geographies

ACS Data Users Conference

May 15, 2019



# Custom geographic areas

- Aggregating tracts or block groups for greater reliability, or fitting natural borders
- Harmonizing geographies over time

Monthly gross rent	Tract A		Tract B		Tracts A + B	
Less than \$250	200	(29%)	0	(0%)	200	(20%)
\$250 to \$499	250	(36%)	50	(17%)	300	(30%)
\$500 to \$999	150	(21%)	150	(50%)	300	(30%)
\$1,000 or more	100	(14%)	100	(33%)	200	(20%)
Total	700	(100%)	300	(100%)	1,000	(100%)

# What about medians?

Monthly gross rent	Tract A	Tract B	Tracts A + B
Less than \$250	200 (29%)	0 (0%)	200 (20%)
\$250 to \$499	250 (36%)	50 (17%)	300 (30%)
\$500 to \$999	150 (21%)	150 (50%)	300 (30%)
\$1,000 or more	100 (14%)	100 (33%)	200 (20%)
Total	700 (100%)	300 (100%)	1,000 (100%)

But the distribution shows that the midpoint is \$499.50!

	Tract A	Tract B	Tracts A + B
Median gross rent	\$450	\$900	????

Users often take a weighted average of the two medians:

$$\frac{(\$450 \times 700) + (\$900 \times 300)}{(700 + 300)} = \$585$$

# Obtaining medians from grouped frequency data

	Frequency	Cumulative frequency	Cumulative percentage
Less than \$300	10	10	10%
\$300 to \$499	10	20	20%
\$500 to \$749	10	30	30%
\$750 to \$999	15	55	55%
\$1,000 to \$1,499	25	80	80%
\$1,500 or more	20	100	100%
Total	100		

30% have rent  $\leq$  \$749  
(\$749 is the 30<sup>th</sup> percentile)

The median (50<sup>th</sup> percentile) is somewhere between \$749 and \$999

55% have rent  $\leq$  \$999  
(\$999 is the 55<sup>th</sup> percentile)

# The formula (back to your first stats class)

	Frequency	Cumulative frequency	Cumulative percentage
Less than \$300	10	10	10%
\$300 to \$499	10	20	20%
\$500 to \$749	10	30	30%
\$750 to \$999	15	55	55%
\$1,000 to \$1,499	25	80	80%
\$1,500 or more	20	100	100%
Total	100		

\$749 = 30<sup>th</sup> percentile

$$+ \left( \left[ \frac{50 - 30}{55 - 30} \right] \times [\$999 - \$749] \right)$$

$$+ ([0.8] \times [\$250])$$

$$+ (\$200) = \$949$$

\$999 = 55<sup>th</sup> percentile

# Why this works

Within the “middle bucket” (\$749 - \$999), we’re assuming a uniform (even) distribution:



- There’s a simple correspondence between percentiles and dollars!
- The 50<sup>th</sup> percentile is 80% of the distance between the 30<sup>th</sup> and 55<sup>th</sup> percentiles, so it’s also 80% of the distance between \$749 and \$999.

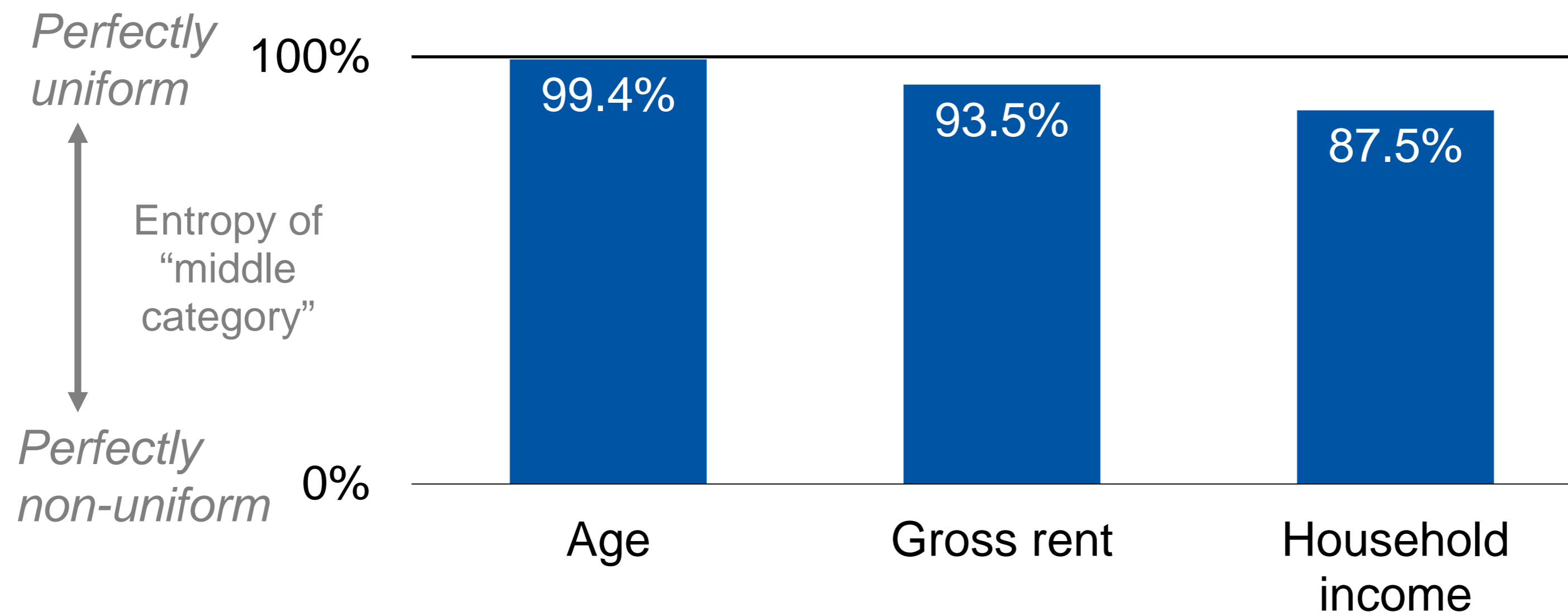
# But what if it doesn't work?

All this rests on the uniformity of the distribution:



# Test 1: Use microdata to examine distributions

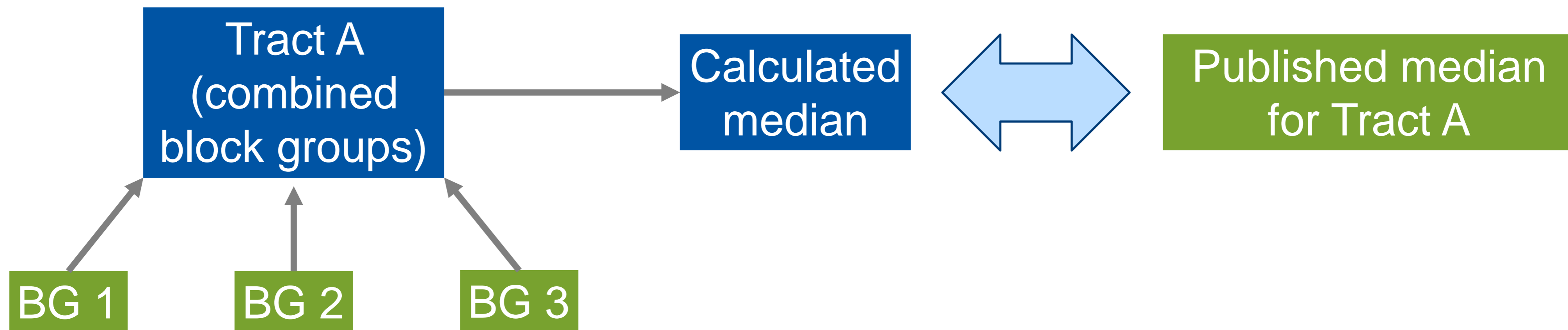
- How uniform are the distributions in these “middle categories”?
- Using the microdata, consult each tract’s PUMA and measure the uniformity of the tract’s “middle category”
- Averaging across all tracts, the “middle category” in their PUMA is fairly uniform (though not so much for household income):





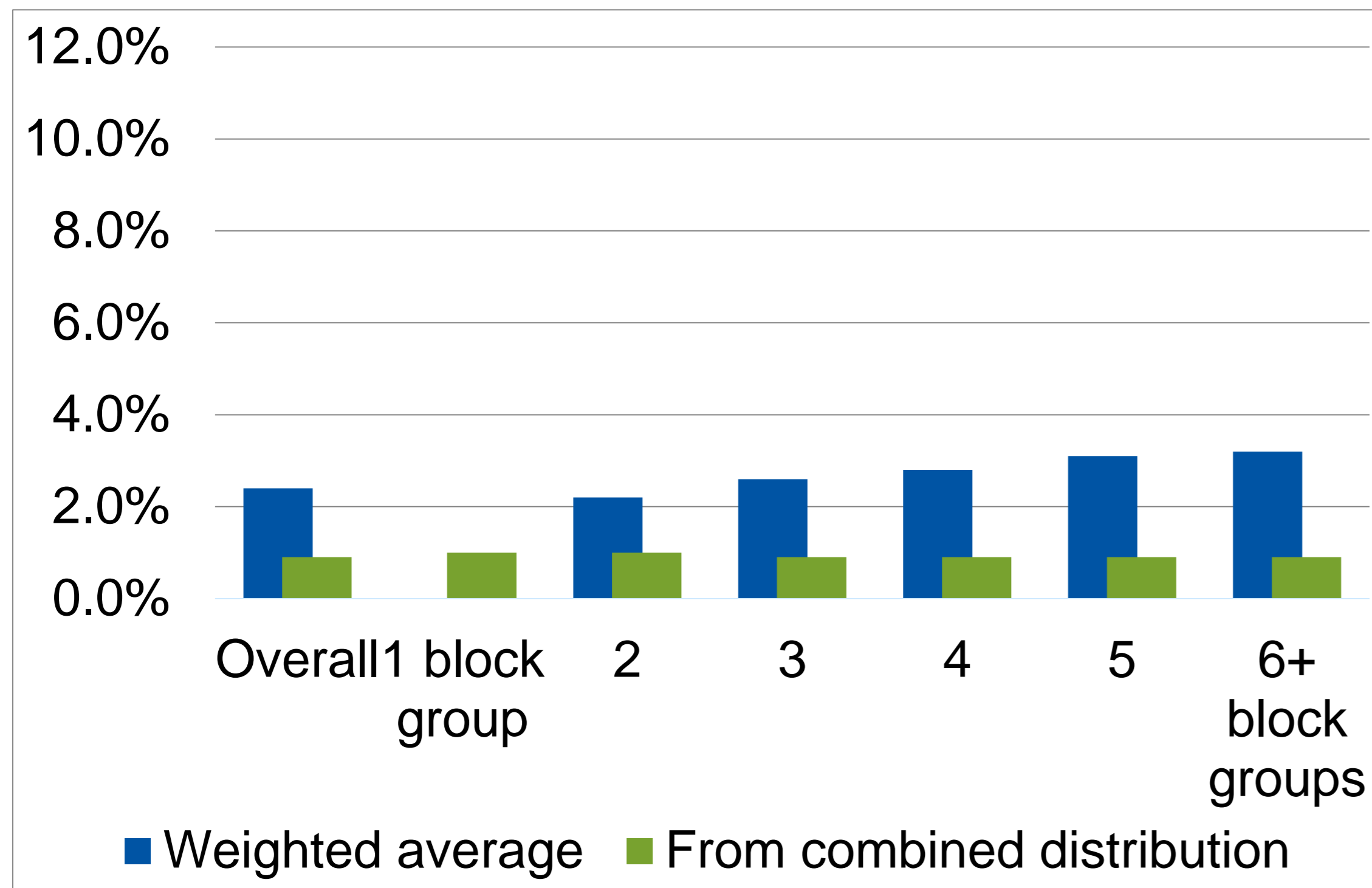
# Test 2: Roll up block groups to tracts

- Sum up the ACS estimates from block groups to tracts
- Calculate the median, assuming a uniform distribution of the “middle category”
- Compare to published tract medians

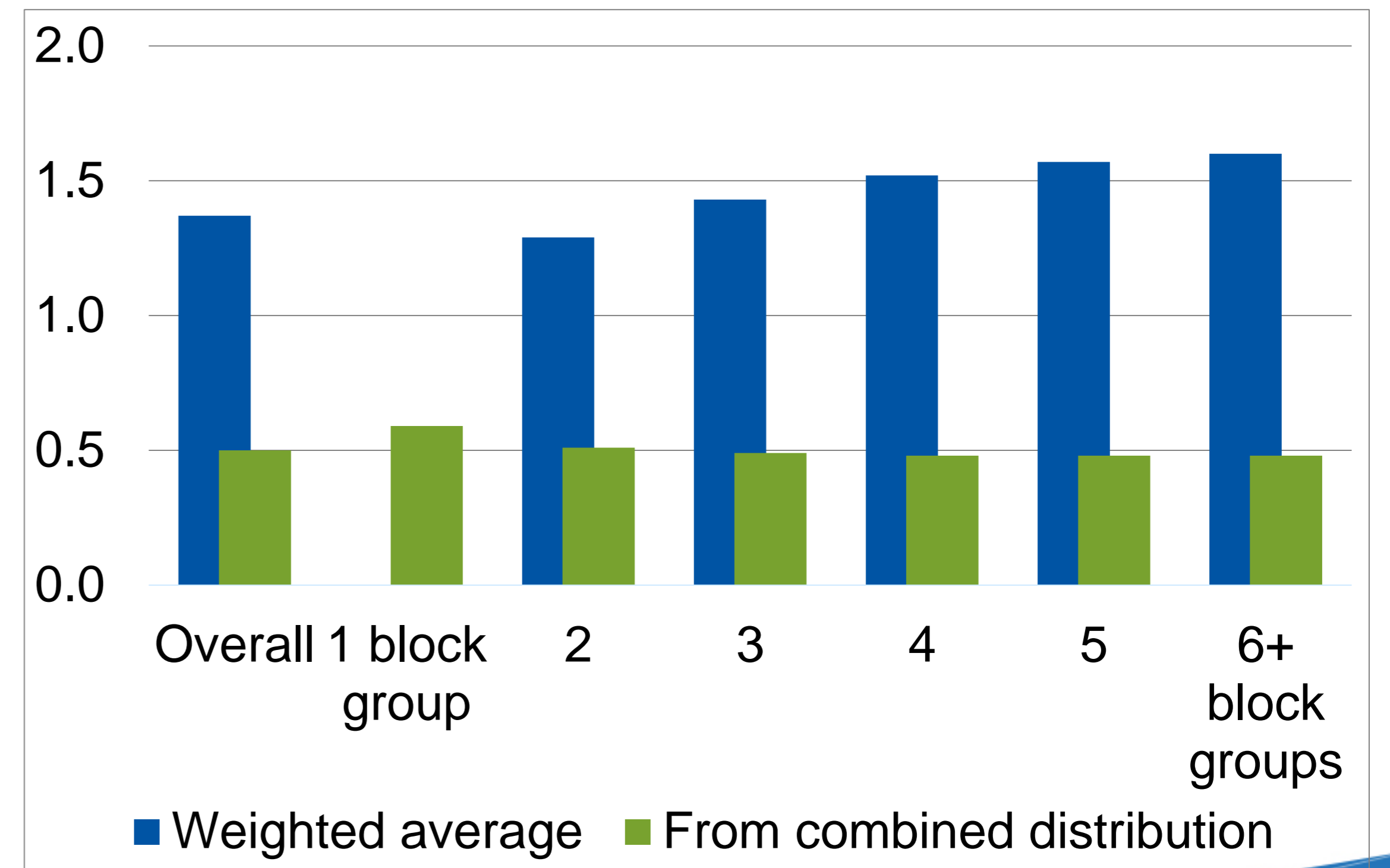


# Test 2: Median age

## Mean percent error

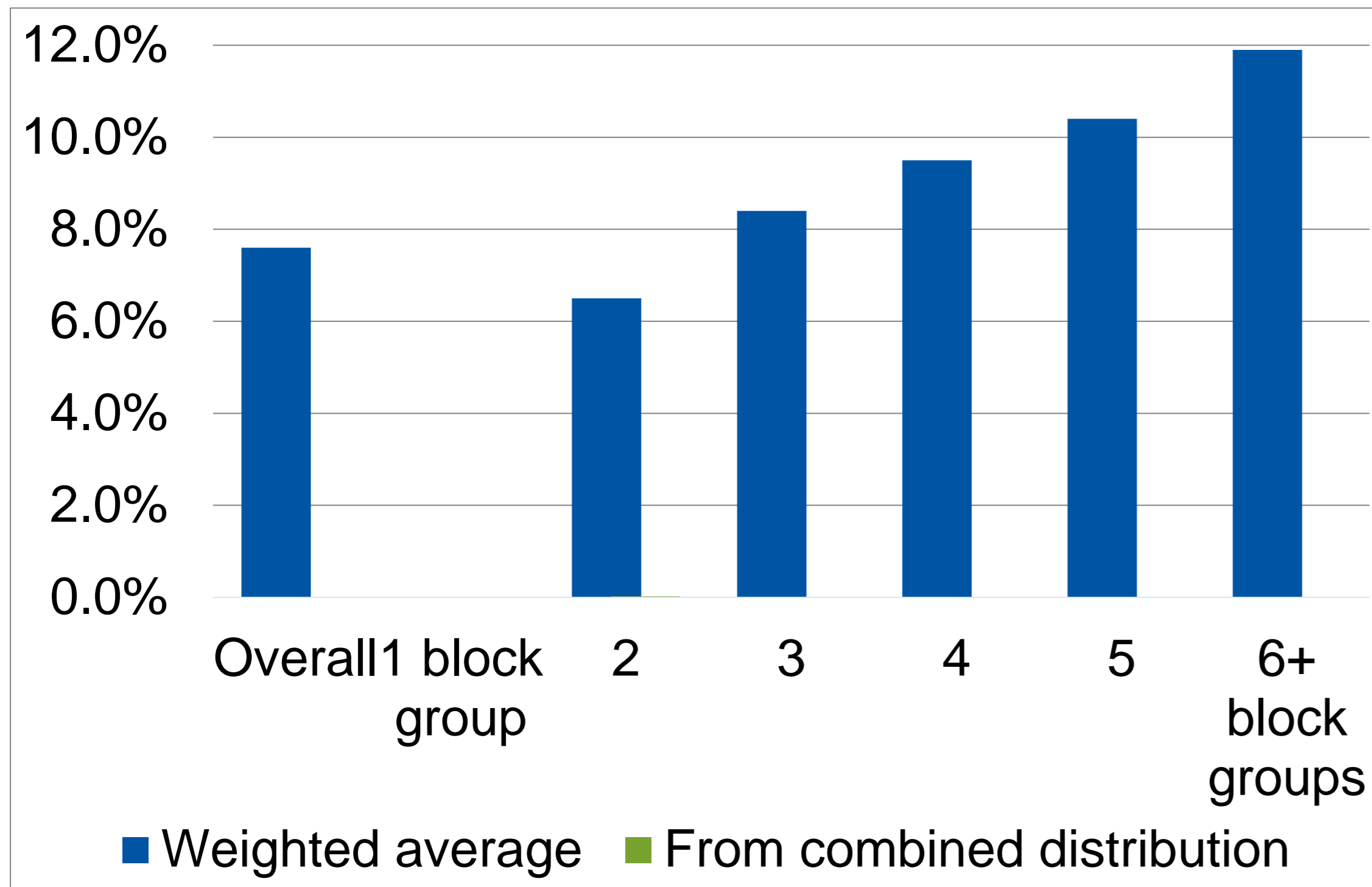


## Root mean square error (in years)

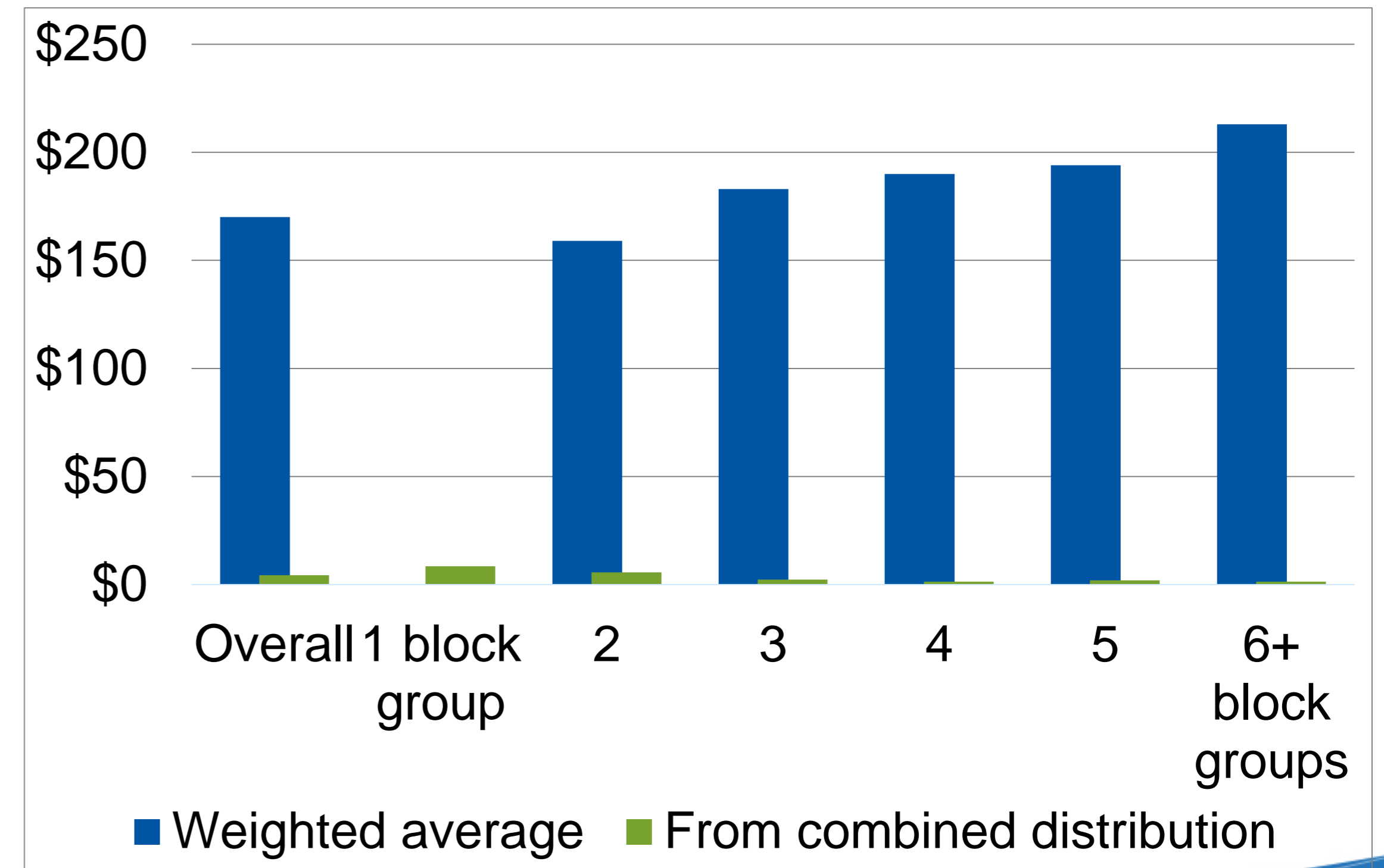


# Test 2: Median gross rent

## Mean percent error

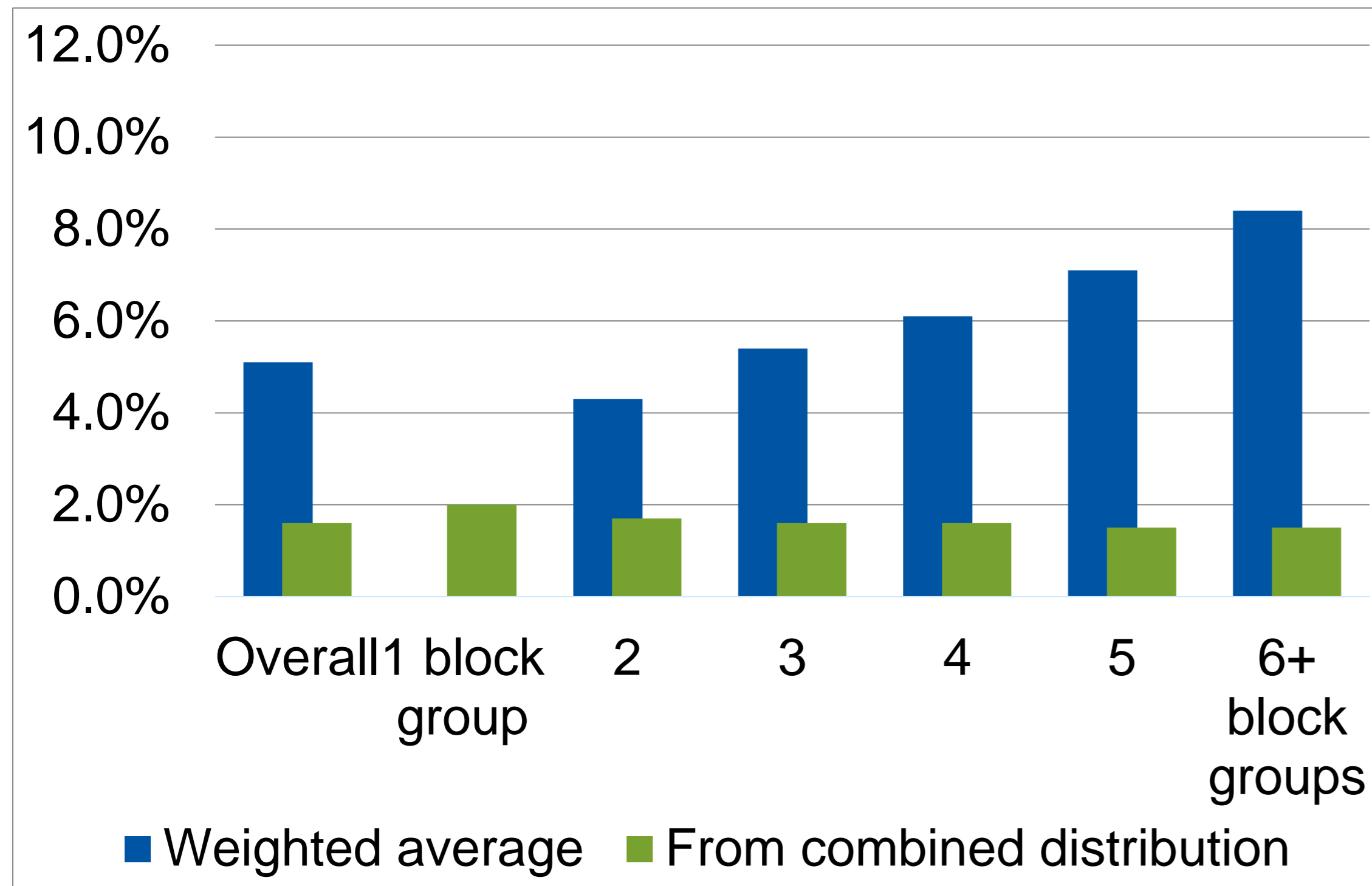


## Root mean square error (in dollars)

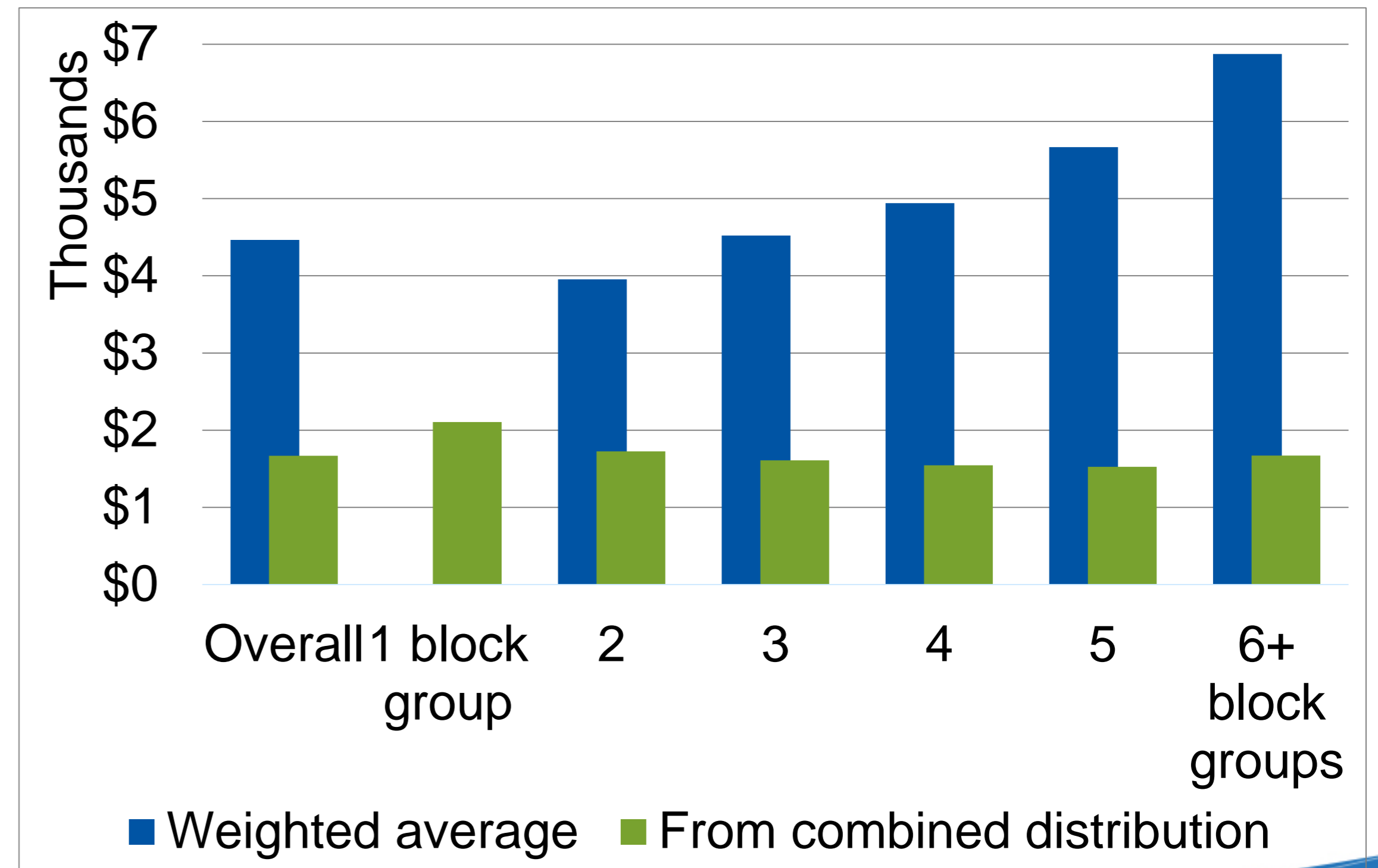


# Test 2: Median household income

## Mean percent error



## Root mean square error (in dollars)



# Summing up

- Estimating the median from the combined frequencies is:
  - Better than taking a weighted average of the medians
  - Pretty accurate overall (<2% average error)
- If increased accuracy is needed, consider simulating the full distribution using the rest of the percentiles and find the median that way

**Questions? Concerns? Want SAS code?**

**[Matt.Schroeder@metc.state.mn.us](mailto:Matt.Schroeder@metc.state.mn.us)**