Assessing the Assessment: Reinterpreting Changes in State- and District-Level NAEP Scores Using a Hierarchical Bayesian Approach

Jennifer Starling, Lauren Forrow, Jon Gellar, and Brian Gill

Federal Conference of Statistical Methodology

October 23, 2024

Release of the 2022 National Assessment of Education Progress (NAEP) showed large nationwide declines in test scores

- Commentators who looked beyond historic nationwide declines in reading and math scores found “bright spots” among places that were “holding steady”
- Reporters saw “no significant change” and thought it meant “holding steady”

The Washington Post

“No change…qualifies as a bright spot”

The New York Times

“In one bright spot, most big-city school districts…held steady in reading”

But the narrative reflected a common misinterpretation of statistical significance

Statistical significance appropriately recognizes the importance of sampling variation

True change for all students in a state or district may differ from the measured change

But statistical significance is not sufficient to tell us whether a result is real or meaningful

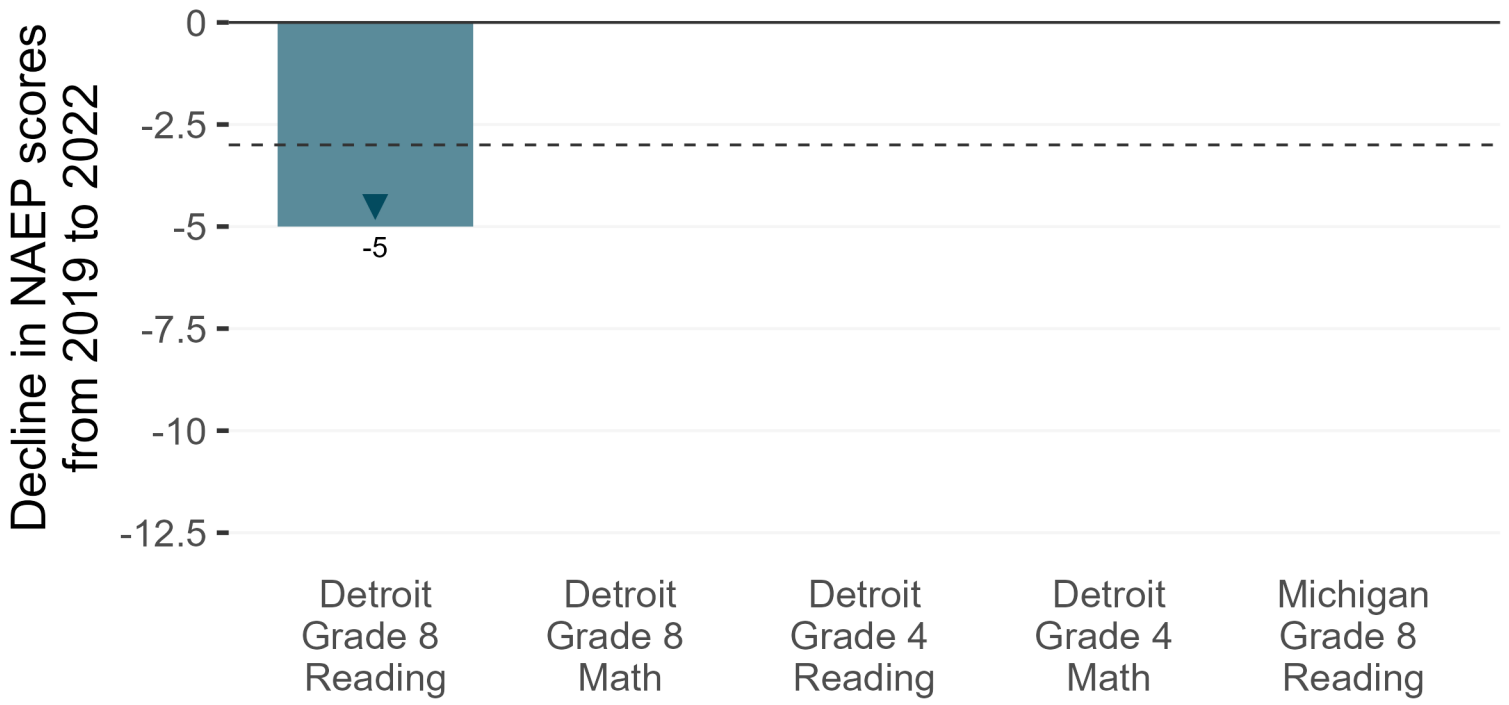
A non-significant result only means:

*Hypothetically, **if** the true score did not change*, the probability that we would see a change this big by chance alone is greater than 5 percent

Absence of a statistically significant result does *not* mean there was no change

A non-significant result can obscure a change large enough to be educationally meaningful

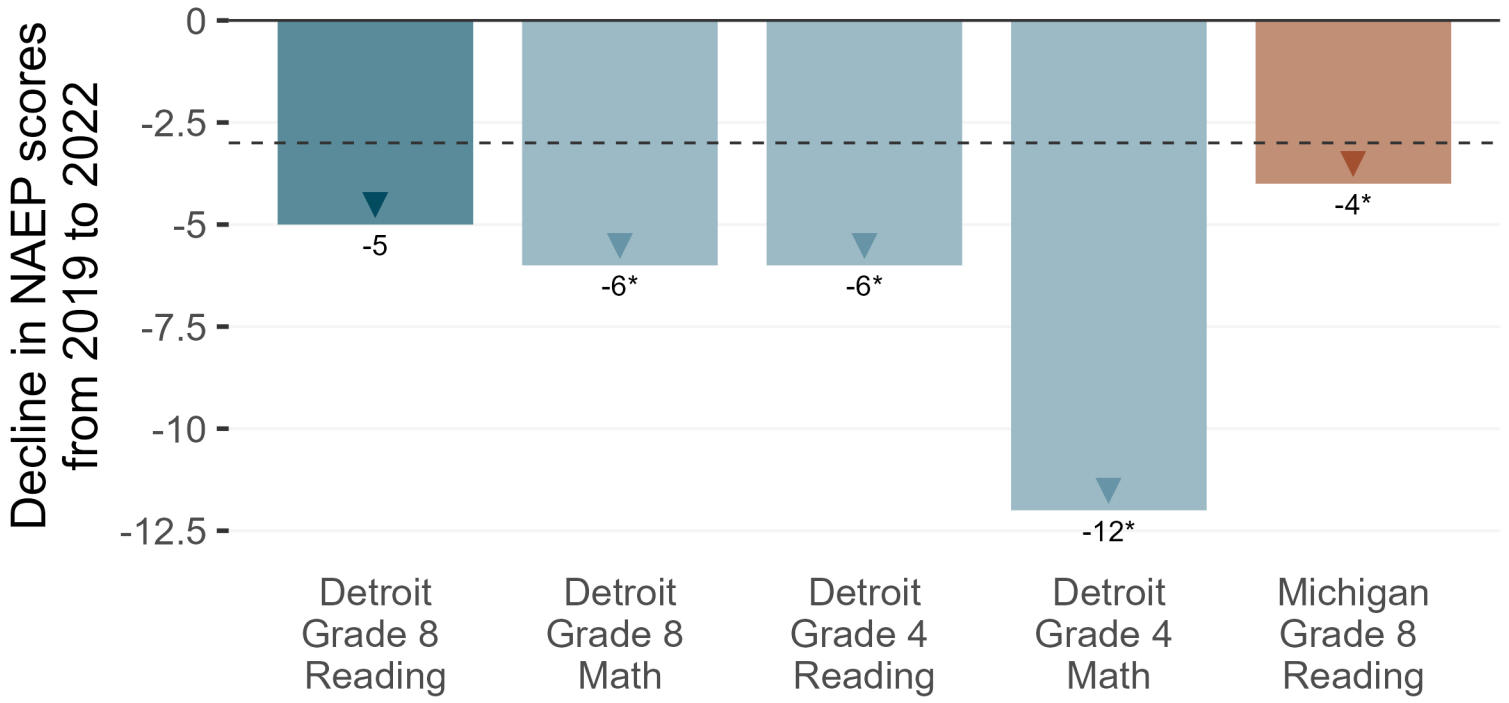
Example: Detroit's 5-point change 8th grade reading scores was not statistically significant



Dashed line: national average decline of 3 in 8th grade reading

Source: NAEP Data Explorer (https://nces.ed.gov/nationsreportcard/data/)

Example: Detroit's 5-point change 8th grade reading scores was not statistically significant

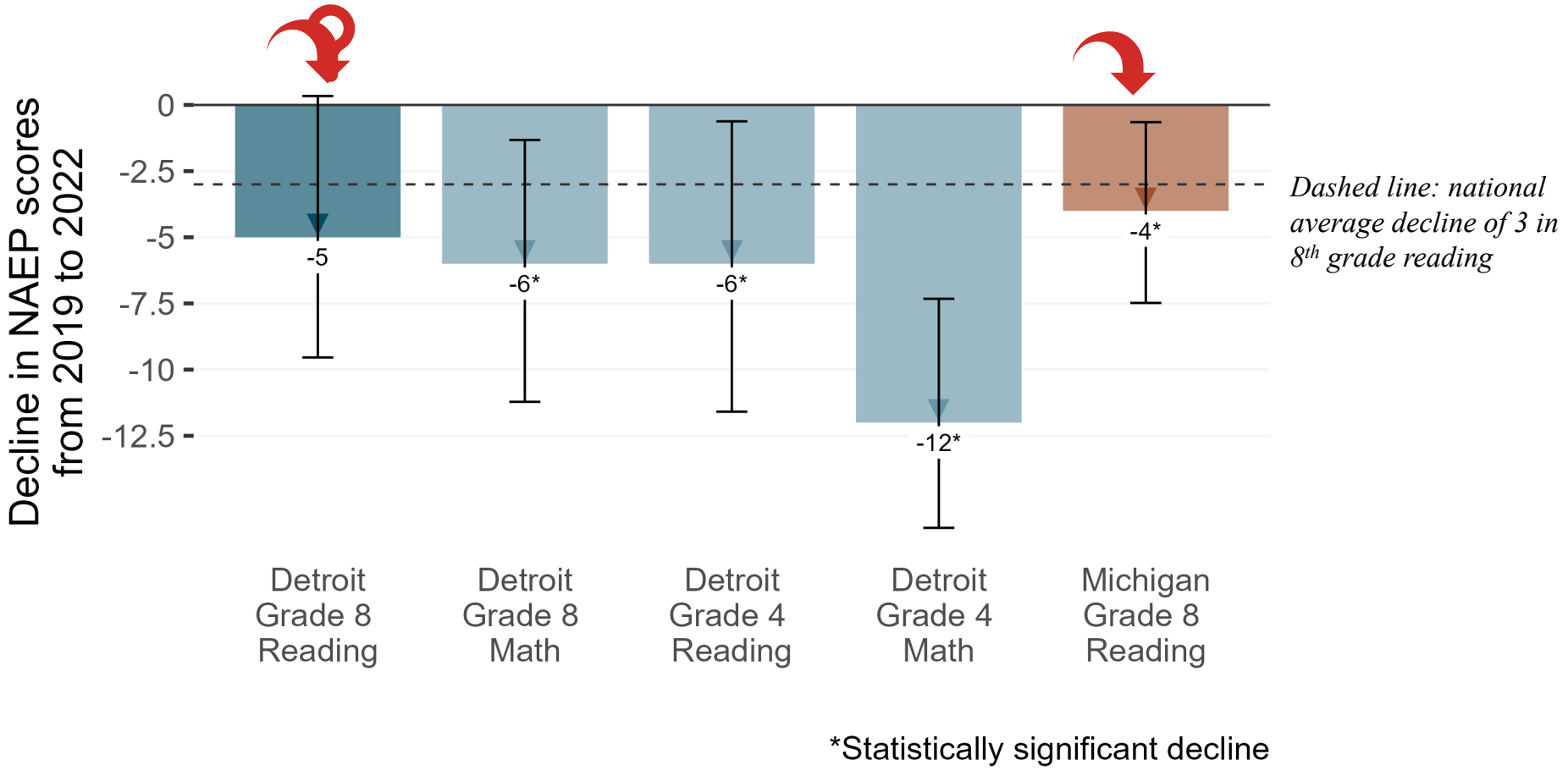


Dashed line: national average decline of 3 in 8th grade reading

*Statistically significant decline

Source: NAEP Data Explorer (https://nces.ed.gov/nationsreportcard/data/)

Example: Detroit's 5-point change 8th grade reading scores was not statistically significant



Source: NAEP Data Explorer (https://nces.ed.gov/nationsreportcard/data/)

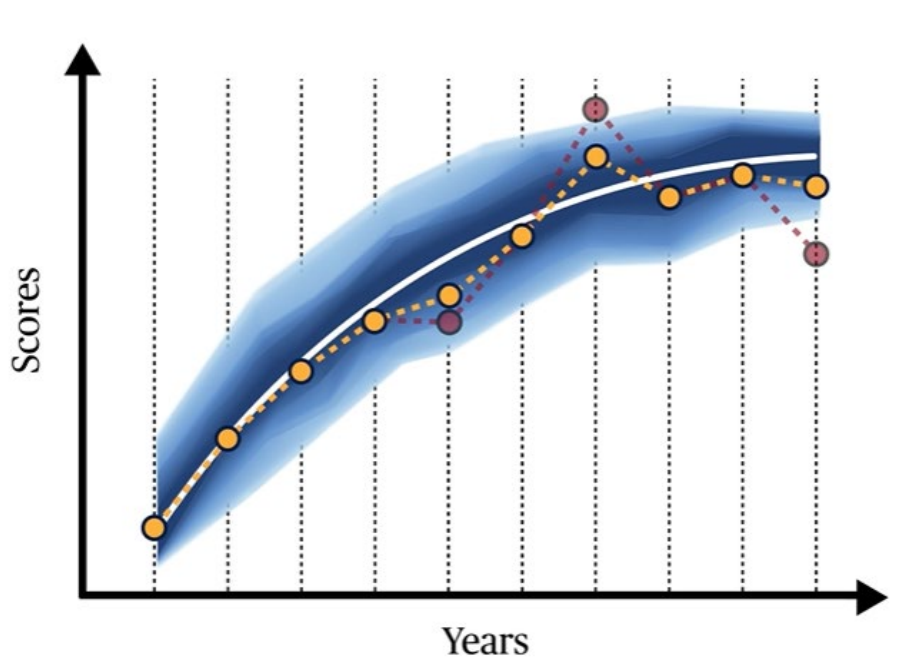
We re-analyzed district and state-level NAEP scores using Bayesian hierarchical modeling

The Bayesian modeling framework provides two benefits:

1. **Bayesian methods stabilize the estimated changes to get a more reliable answer**
   - Adjusts estimates based on context
2. **Bayesian methods directly answer the research question**
   - Delivers more actionable, policy-relevant results
   - For example: “There is an 85 percent chance that District X had a decline in 8th grade reading scores of 3 or more points, an educationally meaningful amount.”

Bayesian modeling uses a rigorous, structured way of borrowing information across data points

- **Borrowing information ("partial pooling") stabilizes estimates toward the overall mean**
  - Estimates with smaller sample sizes borrow more information
  - Outliers borrow more information
- **This process quantifies our intuition**
  - Changes that are consistent with other changes are more likely to be real
  - We are more skeptical of outliers, especially for small sample sizes

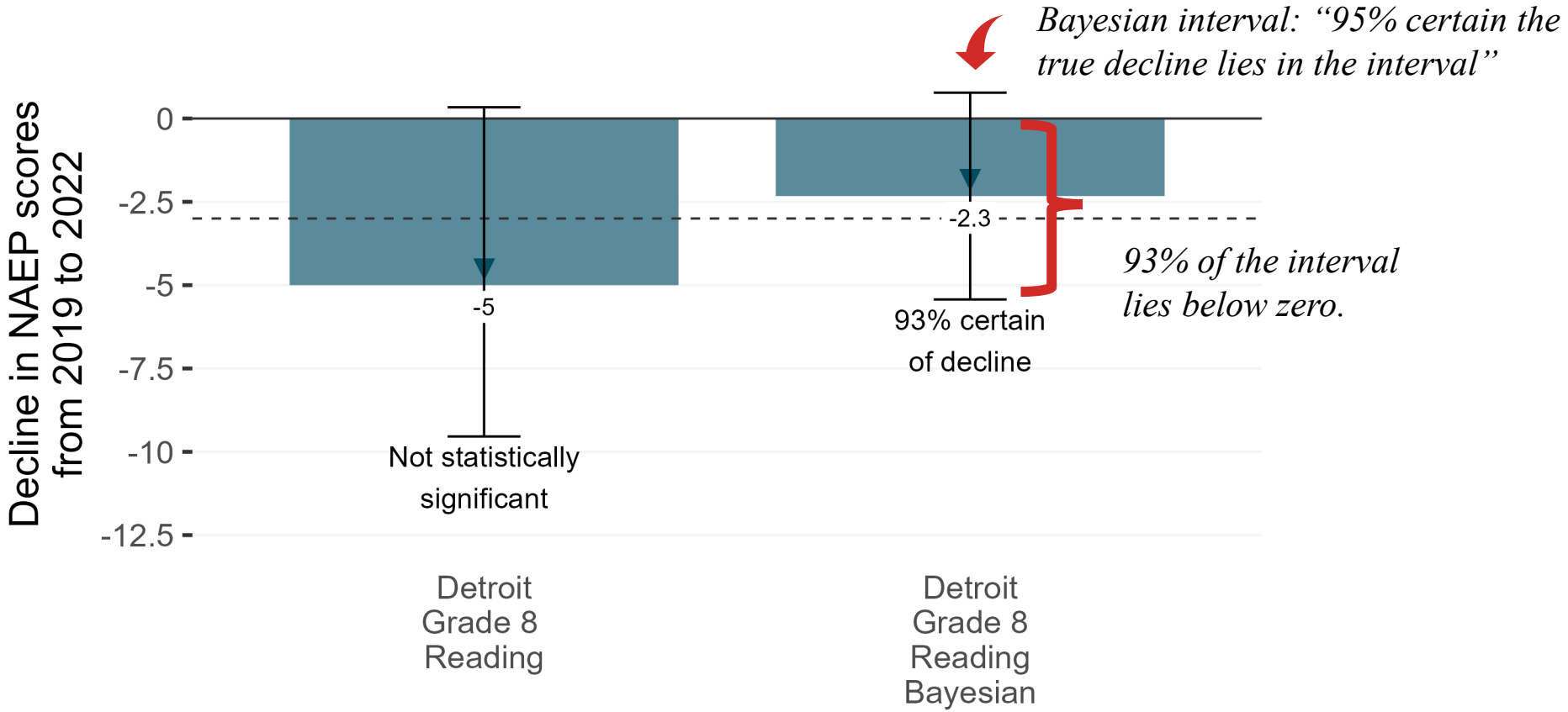


We re-analyzed changes in 4th and 8th grade math and reading scores from 2019 to 2022

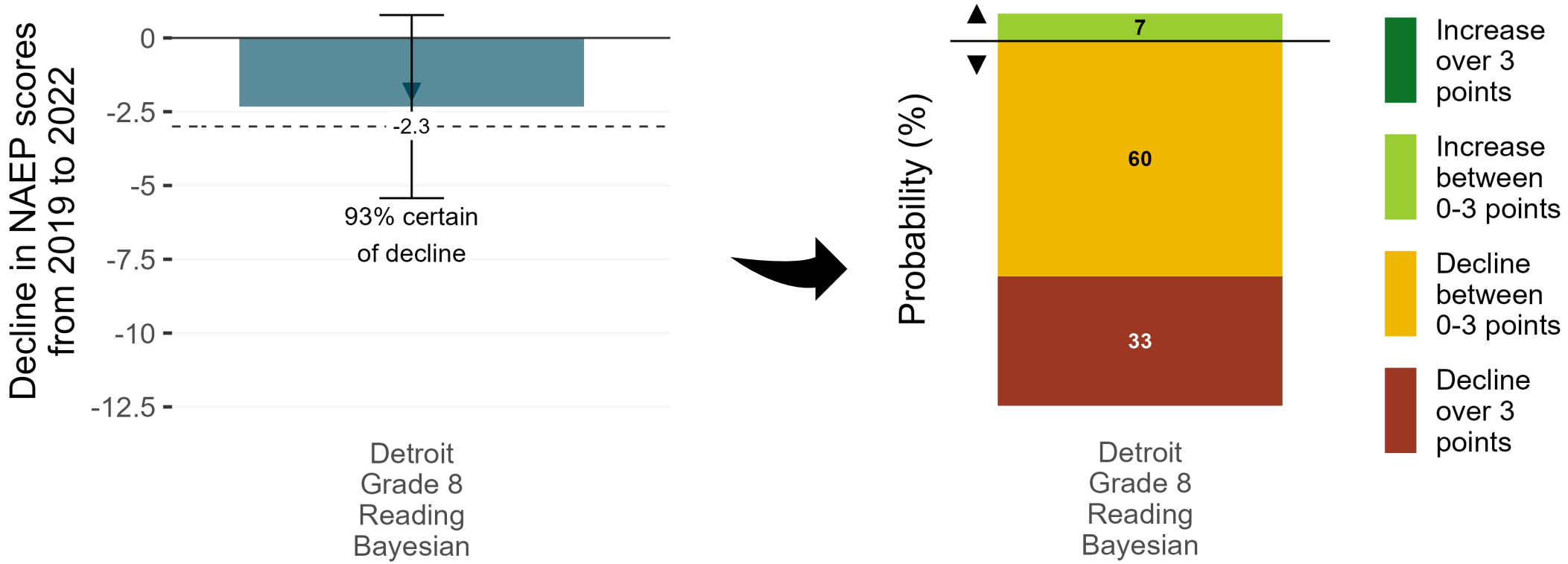
- **Fit two models: One for state- and one for district- level scores**
  - In each model, estimated separated trends for each jurisdiction, grade and subject
- **Each model borrows information across:**
  - Jurisdictions (states or districts)
  - Subjects (math and reading) and grade levels (4th and 8th) within a state or district
- **Used results to describe the magnitude of changes in scores**
  - We describe a decline or increase of 3 or more points as educationally meaningful
  - We describe a decline of 0-3 points as small
  - Jurisdictions are classified according to their most likely scenario for each grade and subject

Results

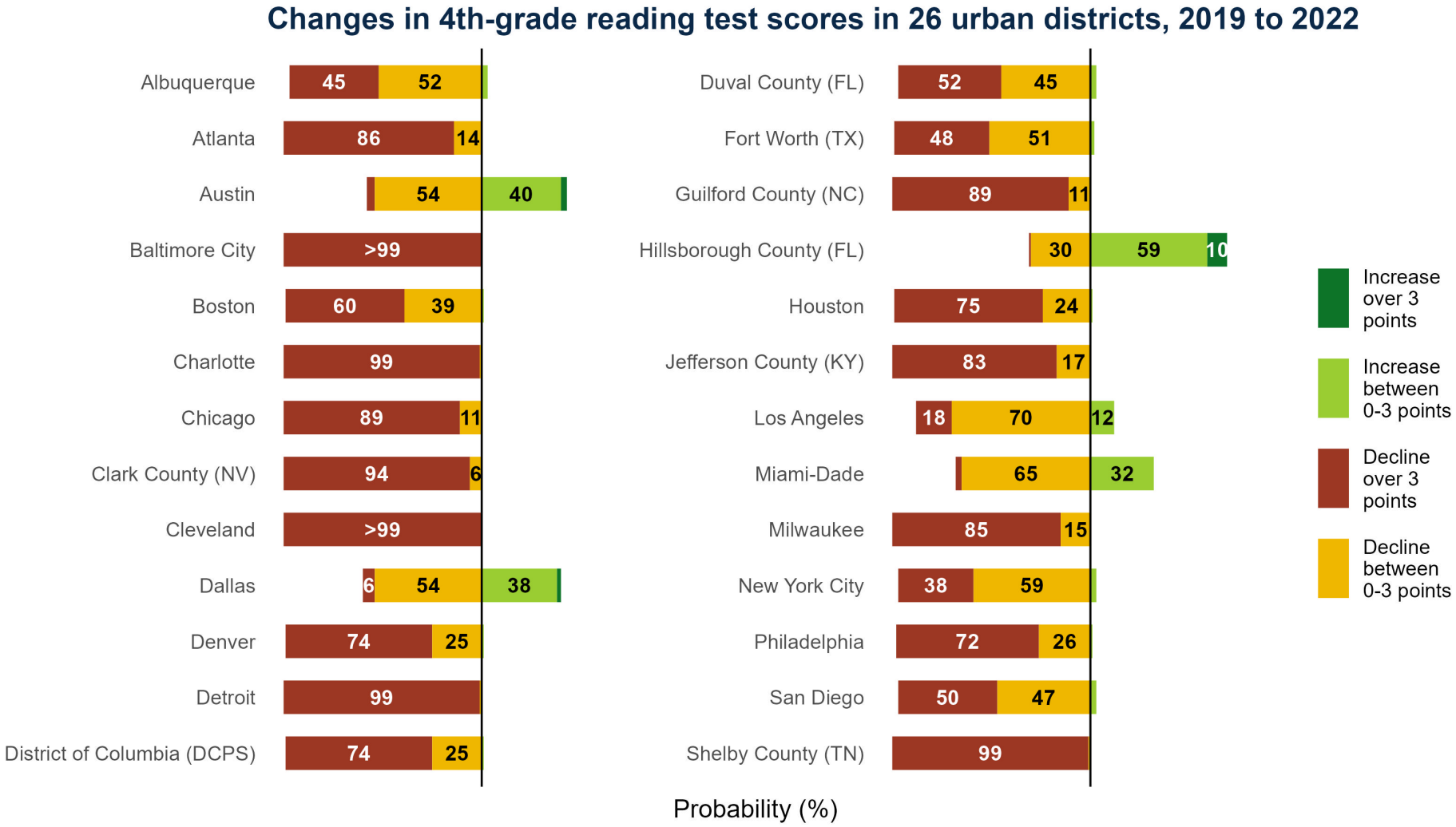
Bayesian results for Detroit's 8th-grade reading scores



Bayesian results for Detroit's 8th-grade reading scores

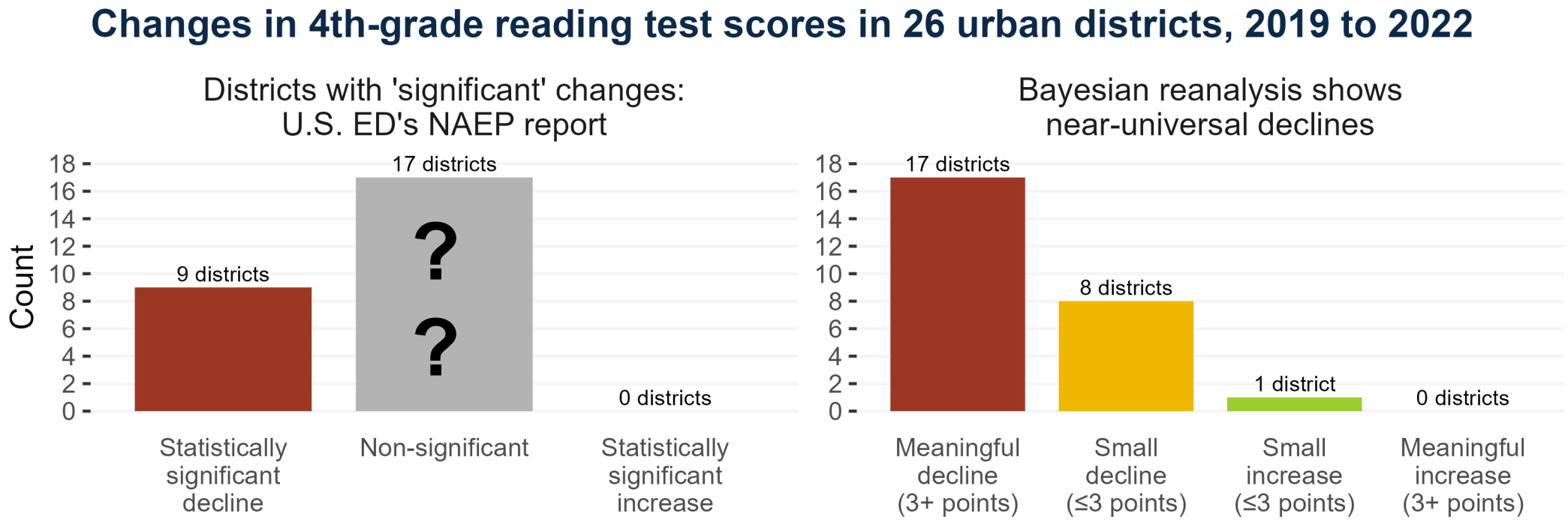


Change in districts' 2019 and 2022 Grade 4 reading scores



*Results for each state, district, subject, and grade at https://www.mathematica.org/blogs/state-and-local-naep-declines-were-more-universal-than-commonly-reported

Bayesian re-analysis shows NAEP declines were nearly universal across districts and states*



In other words: Average scores for fourth-grade reading almost certainly declined in a majority of the participating districts, usually by an educationally meaningful amount

Source: NAEP Data Explorer (https://nces.ed.gov/nationsreportcard/data/)

Conclusions and and opportunities

NAEP can provide richer information about student achievement across the country

- **NCES is right to attend to the risk of misleading flukes**
- **Bayesian analysis deals with random variation in a way that improves the information provided to policymakers and the public**
  - Uses all available information to account for flukes
  - Addresses questions of greatest relevance
  - Avoids presenting results that are easily misinterpreted
  - Provides answers that are intuitively interpretable

Bayesian analysis presents opportunities for better capturing improvement and understanding subgroups

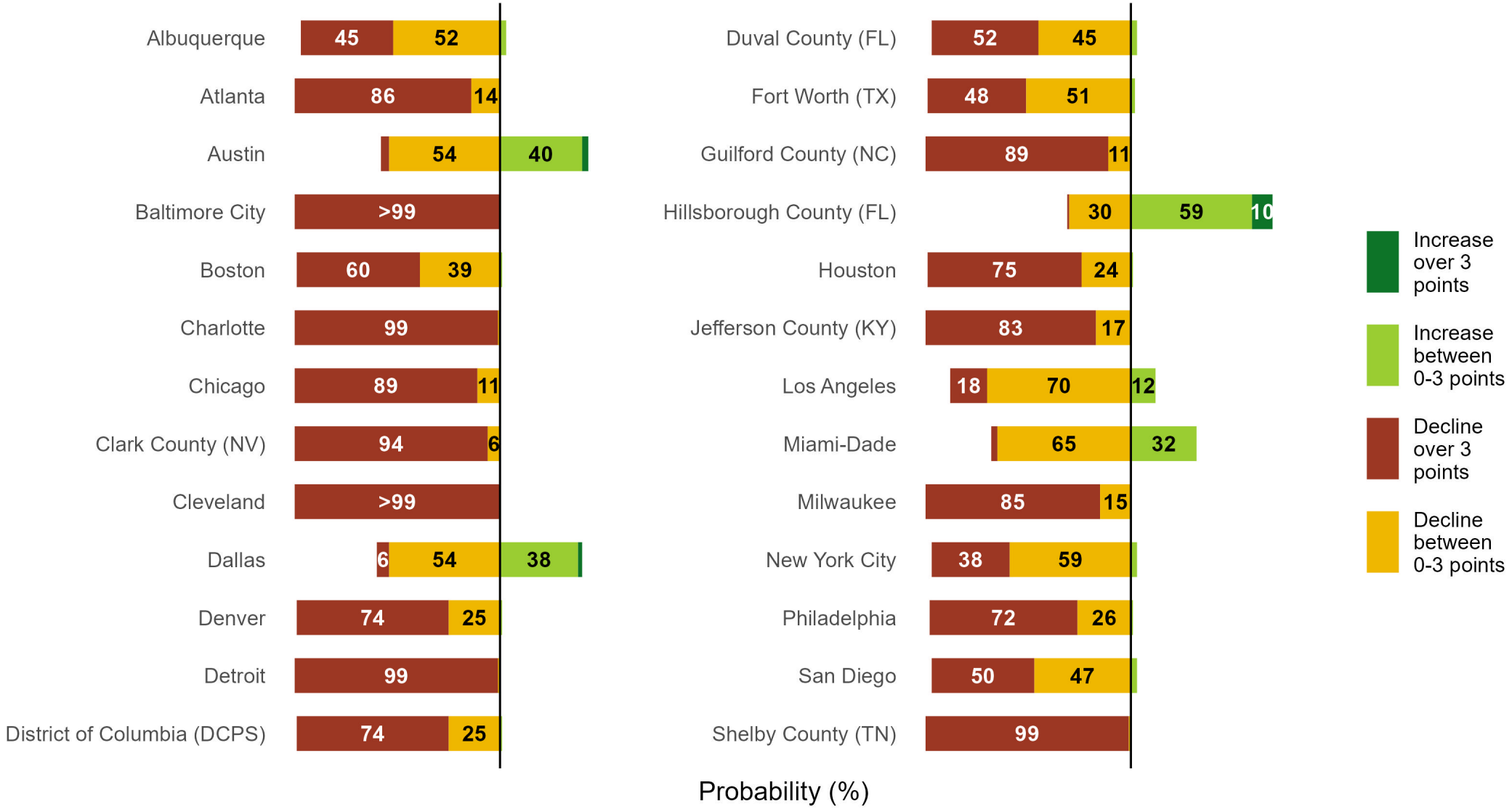
- **If NAEP results substantially improve nationwide, Bayesian analysis will likely show *gains* in more states and districts**
  - Especially helpful for capturing district-level improvements
- **Bayesian analysis can help with understanding subgroup differences and similarities, as well as local results**
  - Increase accuracy and utility of results by race/ethnicity, poverty, disability, region, urbanicity, and more

Thank you!

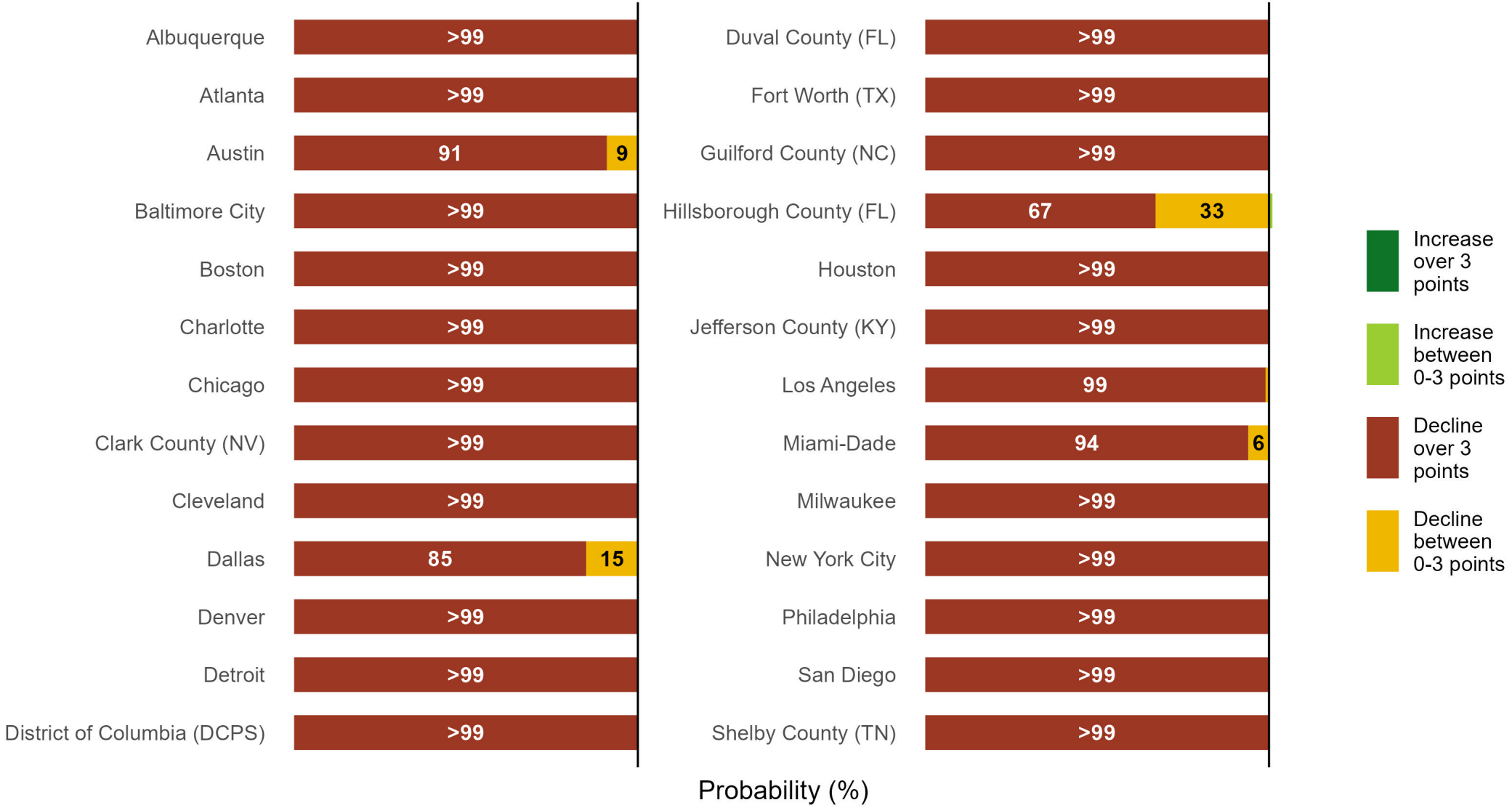
“State and Local NAEP Declines Were More Universal Than Commonly Reported”. Forrow, Starling, Gill, and Gellar. Dec 14, 2022.

Appendix

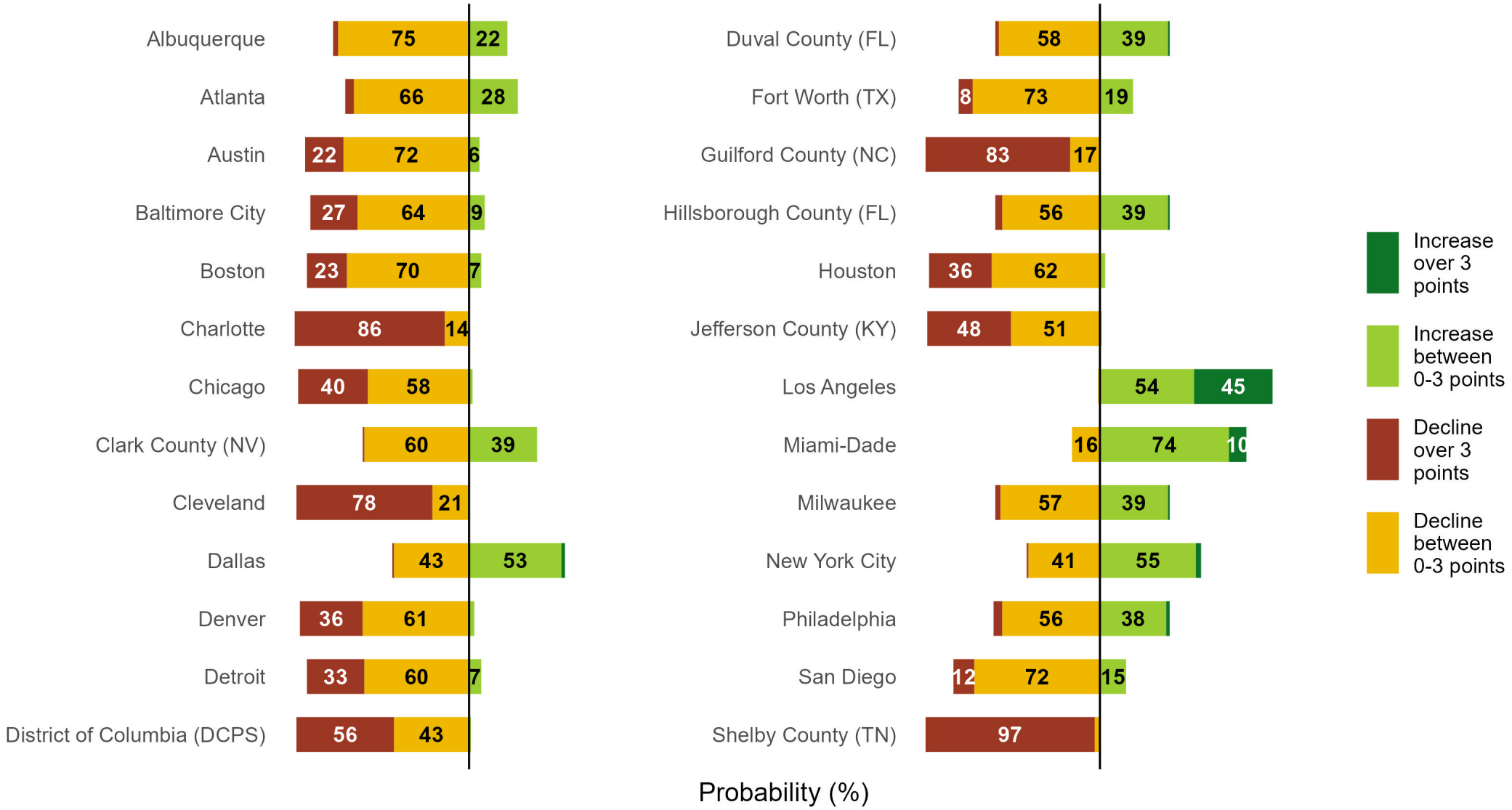
Changes in 4th-grade reading test scores in 26 urban districts, 2019 to 2022



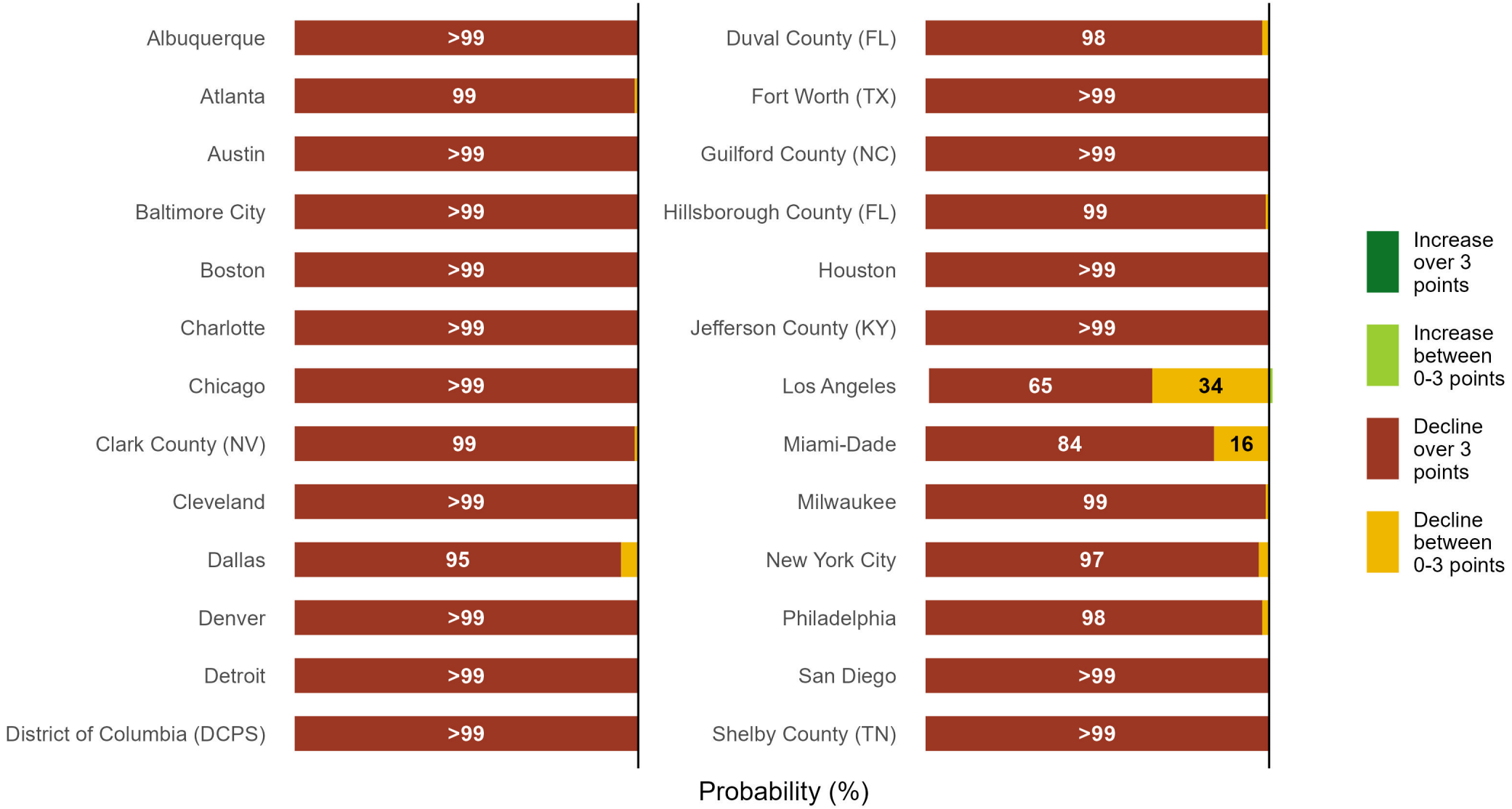
Changes in 4th-grade math test scores in 26 urban districts, 2019 to 2022



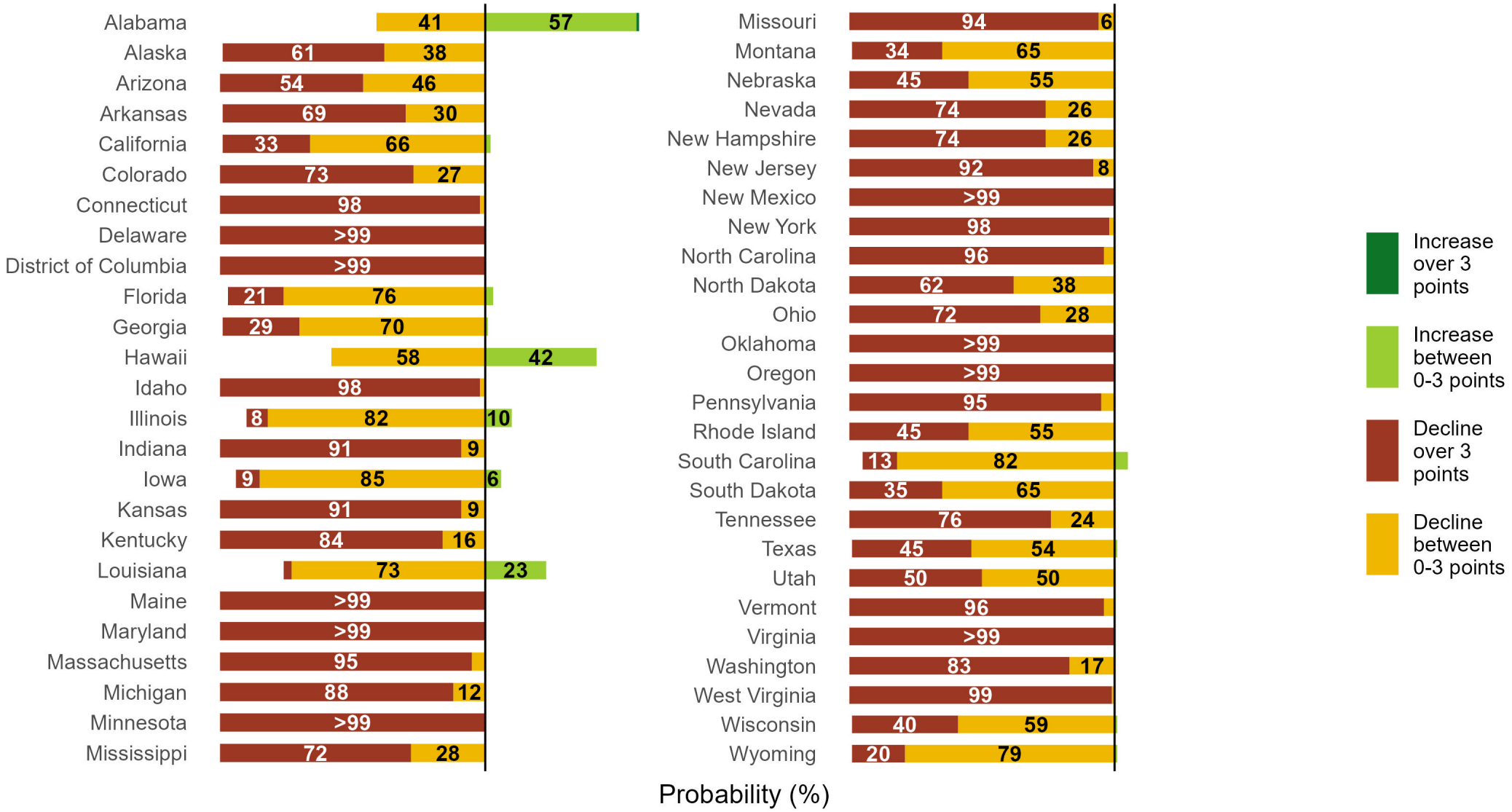
Changes in 8th-grade reading test scores in 26 urban districts, 2019 to 2022



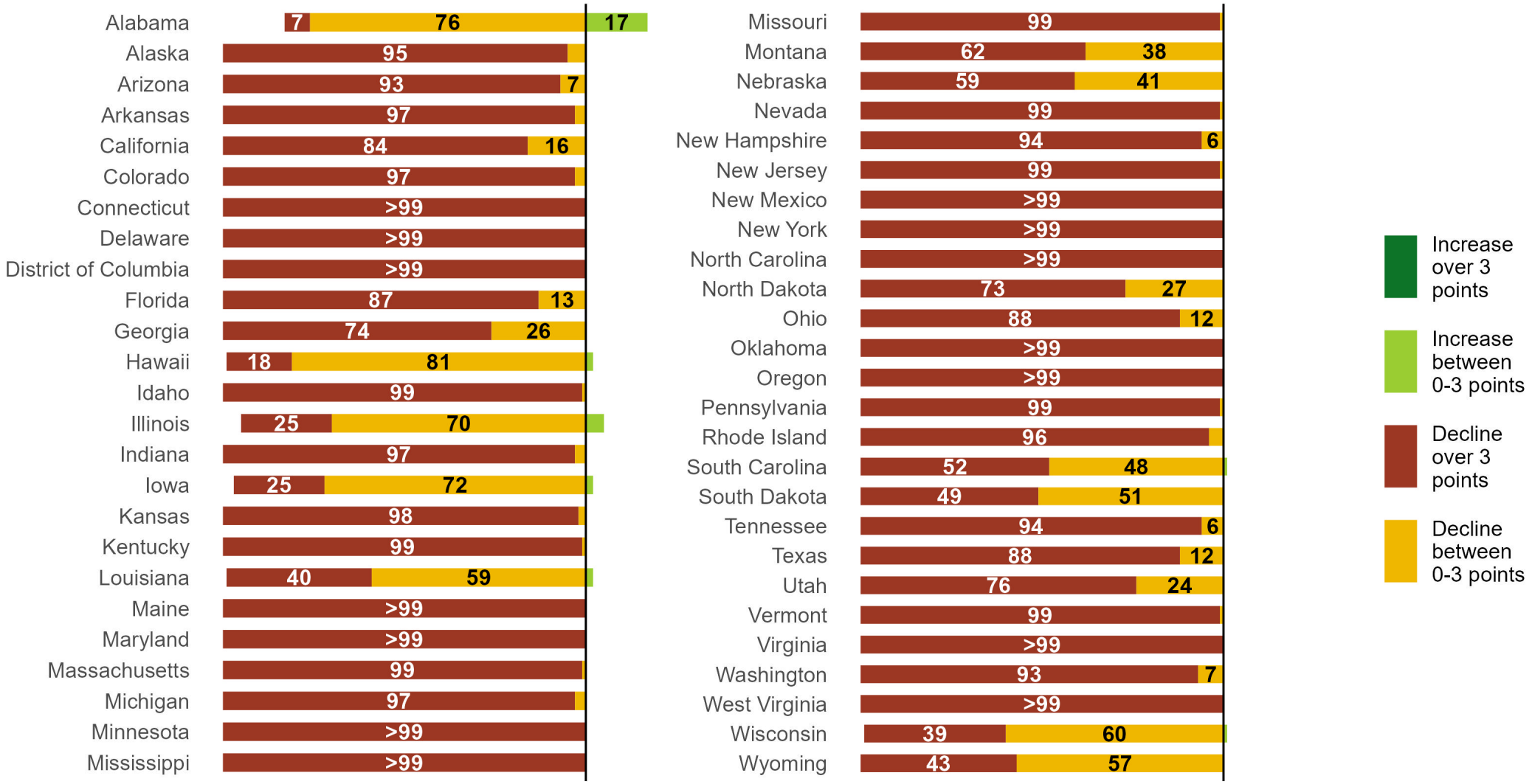
Changes in 8th-grade math test scores in 26 urban districts, 2019 to 2022



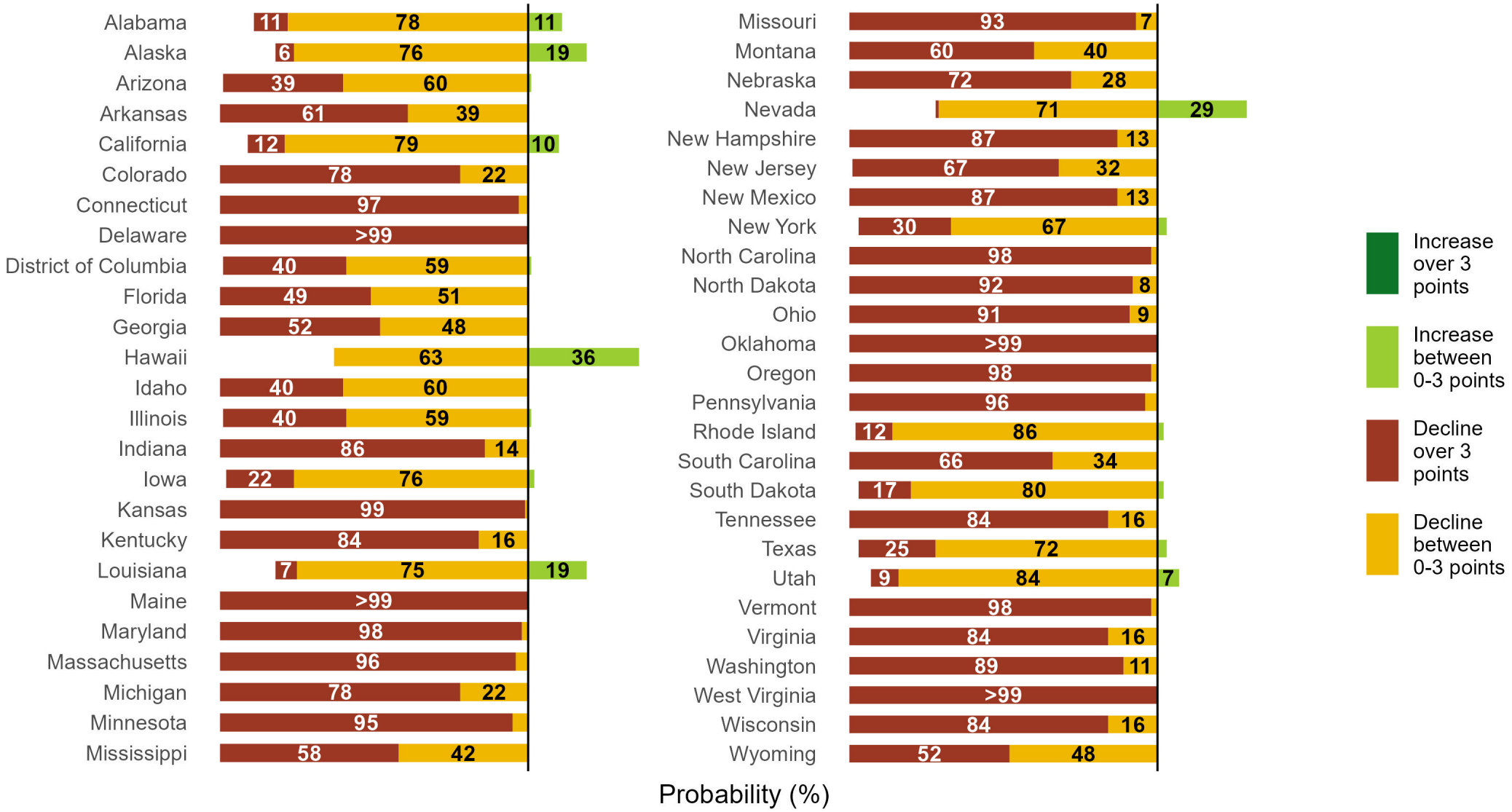
Changes in 4th-grade reading test scores by state, 2019 to 2022



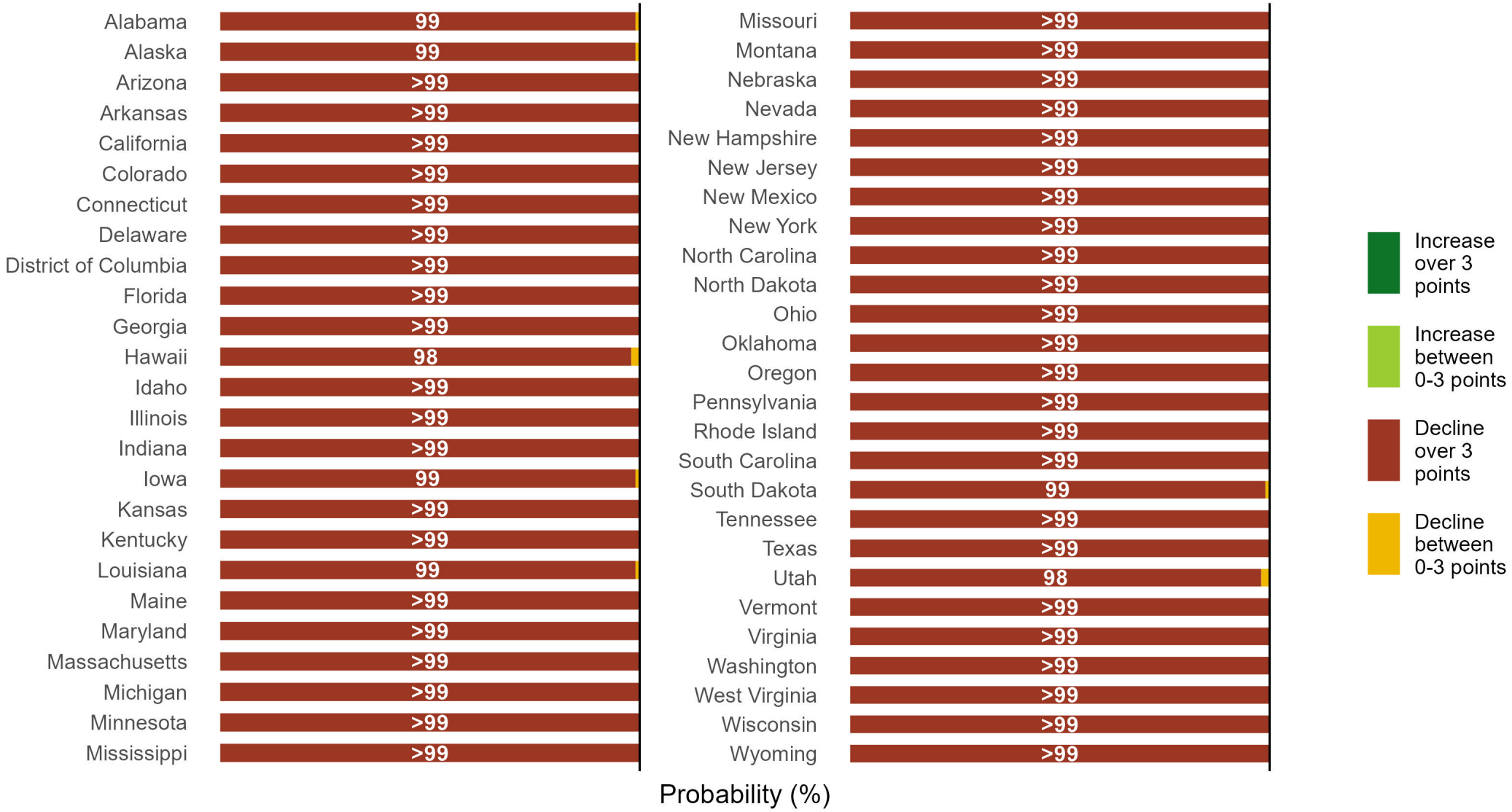
Changes in 4th-grade math test scores by state, 2019 to 2022



Changes in 8th-grade reading test scores by state, 2019 to 2022



Changes in 8th-grade math test scores by state, 2019 to 2022



Methods

Our re-analysis fit two Bayesian models - one for states, and another for districts - that borrow strength across subjects, grades, and jurisdictions. Conforming with best practices in the literature, we chose weakly informative prior distributions that assume that parameters governing variability should not be too large.

We fit our models using Hamiltonian Monte Carlo as implemented in the Stan probabilistic programming language and assessed convergence and mixing using the Gelman-Rubin diagnostic and effective sample sizes.

Our models used imputed 2019 scores for Los Angeles, as Los Angeles excluded charter schools on a one-time basis in 2019 (which comprise nearly 20% of Los Angeles' public schools).

Model specification

We write each of our Bayesian models as follows, where jurisdictions represent states or districts, respectively. Let j represent jurisdictions, s represent subject (Math or Reading), and g represent grade (fourth or eighth). Let t indicate academic year (2018/19 or 2021/22).

Then y_{jtsg} gives the NAEP score for jurisdiction j in year t for subject s in grade g.

$$
\begin{aligned}
y_{jtsg} &= \alpha_{jsg} + \delta_{jsg} I_{\{t=2022\}} + \epsilon_{jtsg} \\
\alpha_{jsg} &= \alpha_j^0 + \alpha_j^S S_s + \alpha_j^G G_g + \alpha_j^X S_s G_g \\
\delta_{jsg} &= \delta_j^0 + \delta_j^S S_s + \delta_j^G G_g + \delta_j^X S_s G_g \\
\epsilon_{jtsg} &\sim N(0, \sigma^2_{jtsg})
\end{aligned}
$$

where standard errors σ_{jtsg} are specified using values from the NAEP data. In this parametrization, we let

- $S_{Reading} = -0.5$
- $S_{Math} = 0.5$
- $G_4 = -0.5$
- $G_8 = 0.5$

so that neither grade or subject is considered a baseline value. (Note that under this parametrization, the α_j^0 and δ_j^0 terms do not refer to a specific grade or subject, so are not directly interpretable.)

The eight random effects (four α_{jsg}'s, and four δ_{jsg}'s, for each subject-grade combination) are assigned prior distribution $MVN(\theta_0, \Sigma)$, with an LKJ prior on Σ. We transform the NAEP scores to z-scores prior to fitting the model and assign other parameters standard normal priors, reflecting a gentle assumption that these parameters are unlikely to be too large.

Model fitting and validation

We fit our model using Hamiltonian Monte Carlo as implemented in the Stan probablistic programming language (Stan Development Team, 2021), via its R interface, rstan. Specifically, we used the brms R package to implement our model using rstan.

We specified our brms model statement as follows, where y represented NAEP scores, y_se represented the corresponding standard errors, Y2022 is an indicator for the 2021/22 academic year, and grade_ctr and subj_ctr represent the *Ss* and *Gg* variables defined previously.

$$y \mid se(y\_se) \sim Y2022 * grade\_ctr * subj\_ctr + (1 + Y2022 * grade\_ctr * subj\_ctr \mid jurisdiction)$$

We assessed convergence and mixing using the Gelman-Rubin diagnostic and effective sample sizes.

- For both our local and state models, Gelman-Rubin statistics were well within recommended ranges for all parameters (from 0.99 to 1.01 for both models).
- Effective sample sizes for all parameters were sufficient, with minimums of 838 for the local model and 506 for the state model.

Imputed scores for Los Angeles

Prior to fitting our models, we imputed two values for each subject-grade combination for Los Angeles in 2019 – the NAEP score, and its standard error.

- We imputed Los Angeles' scores by calculating the percentile across districts that Los Angeles achieved in 2017 and assigning the corresponding 2019 percentile, separately by grade and subject.
- We used the same approach for standard errors, calculating the percentile of standard errors across districts for Los Angeles in 2017, ensuring that both the score itself and the level of precision reflect realistic scenarios based on Los Angeles' 2017 performance.