Improved Inference in Periodic Testing with State-space Models and Growth Estimates

Carson Cook, Garron Gianopulos, Emily Lubkert, Yeow Meng Thum, and S. Thomas Christie

nwea

Introduction

Summative scores for through-year tests

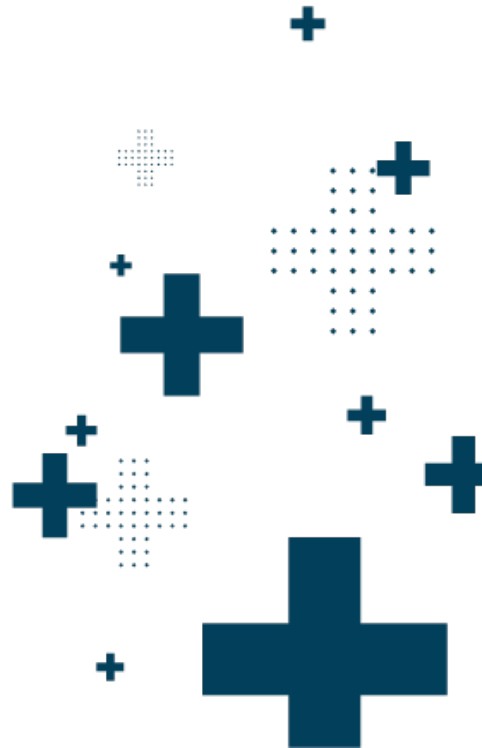
Two forms of assessment have historically dominated in US

+ Summative: a single large test at the end of the year
+ Interim: multiple small tests throughout the year

Through-year assessments seek to blend these approaches

+ Multiple interim tests
+ Final, summative score

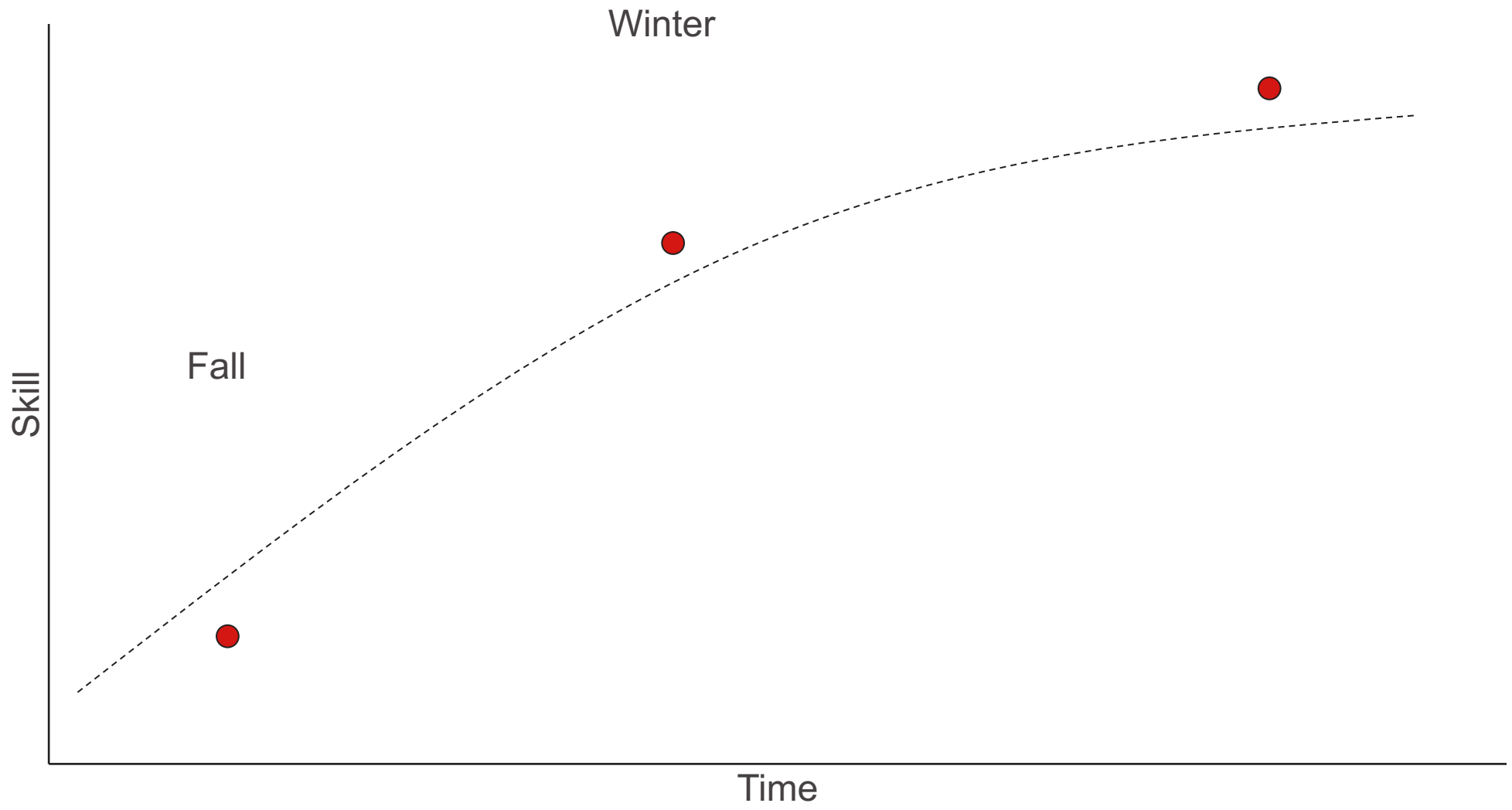
For through-year tests, how do we arrive at a summative score?



nwea

Background

Problem setting



nwea

Existing approaches for summative scores

Simple methods

- Most recent
- Max
- Simple average
- Unweighted/weighted sum

(Wise, 2011)

Bayesian updates

- Custom Bayesian latent trait models with hand-tuned parameters + predictive growth model

(Van Moere, 2020)

Other approaches

- Market basket models that predict performance on reference test + predictive growth model

(Zwick and Mislevy, 2011)

Bayesian state space modeling

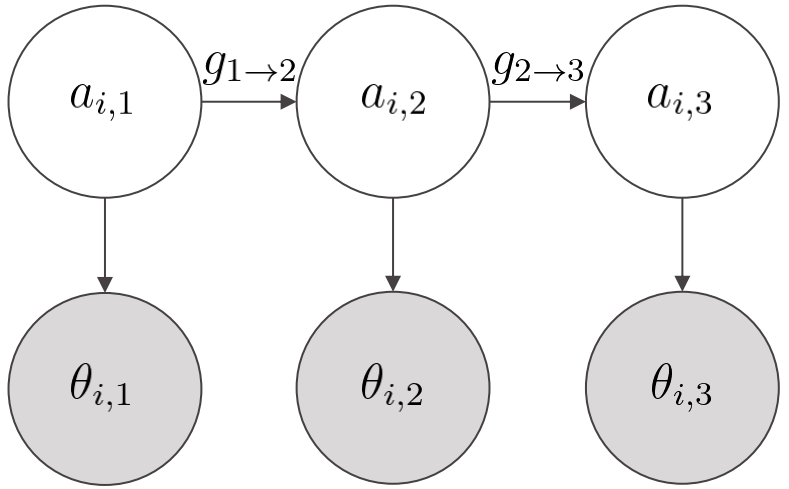
Model a latent student parameter through time based on noisy measurements:

+ Measurement θ typically assumed to follow

$$\theta_t \sim \mathcal{N}\left(a_t, \nu_t\right)$$

+ Latent variable a typically assumed to follow

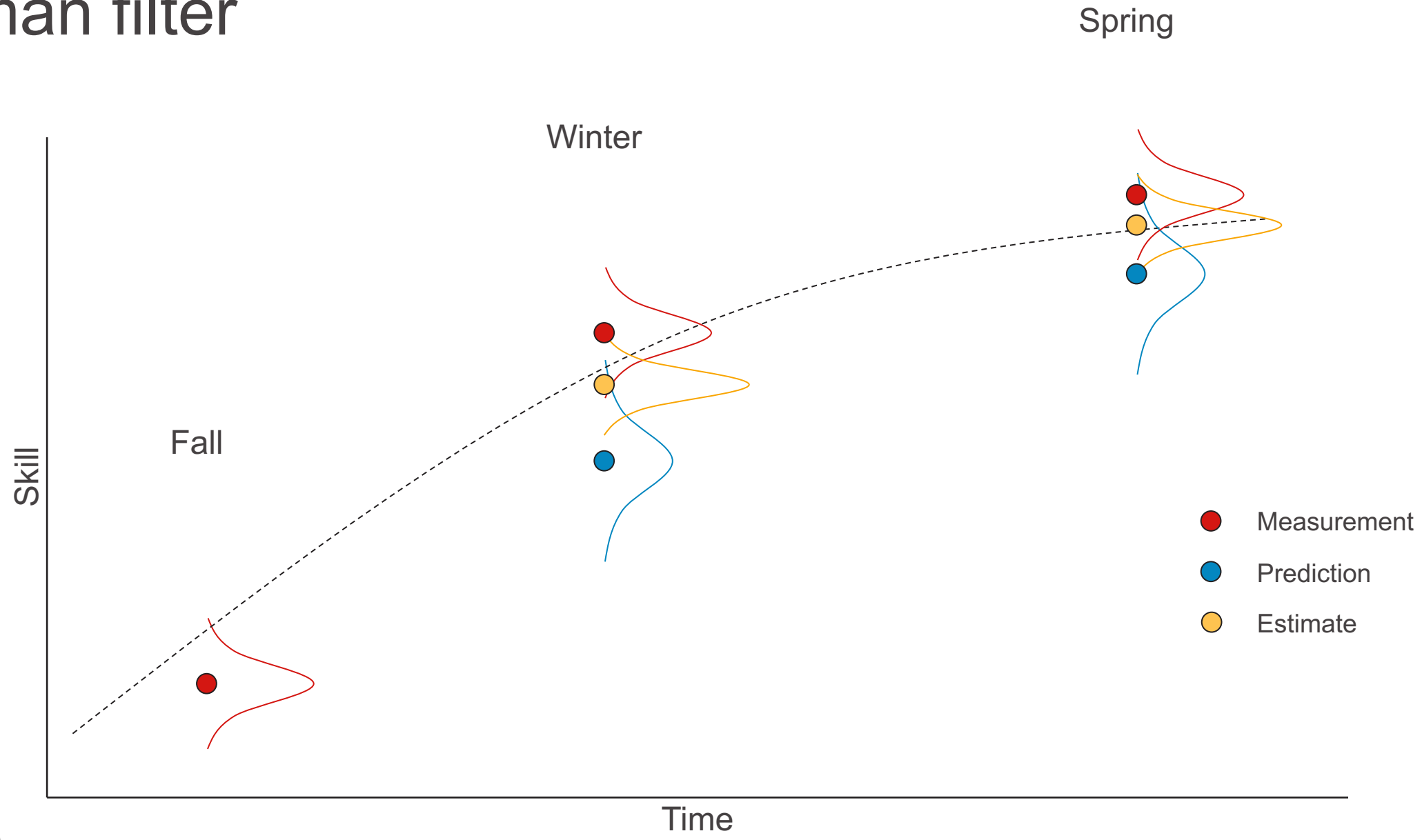
$$a_t \sim \mathcal{N}\left(a_{t-1} + g(a_{t-1}, \Delta t), \gamma_t\right)$$



Here, ν describes measurement error, g describes expected growth of a, and γ describes growth variability across students.

nwea

Kalman filter



nwea

Kalman filter (cont'd)

Analytical method to generate estimates by combining measurements and predictions.

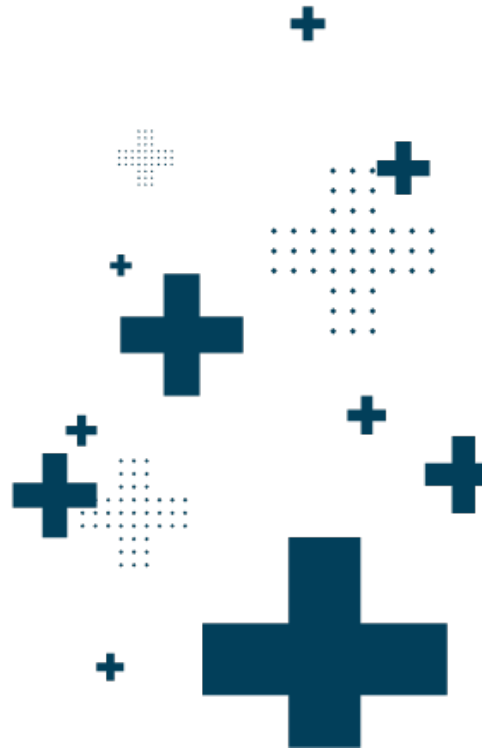
Assumes:

+ Measurement and latent variable normally distributed at each time point
+ Measurement and process noise known at each time point
+ Growth function is a linear function of the state of the system

Estimate update function:

$$\hat{a}_t = \hat{a}_{t-1} + g(\hat{a}_{t-1}, \Delta t) + K_n \left(\theta_t - (\hat{a}_{t-1} + g(\hat{a}_{t-1}, \Delta t))\right)$$

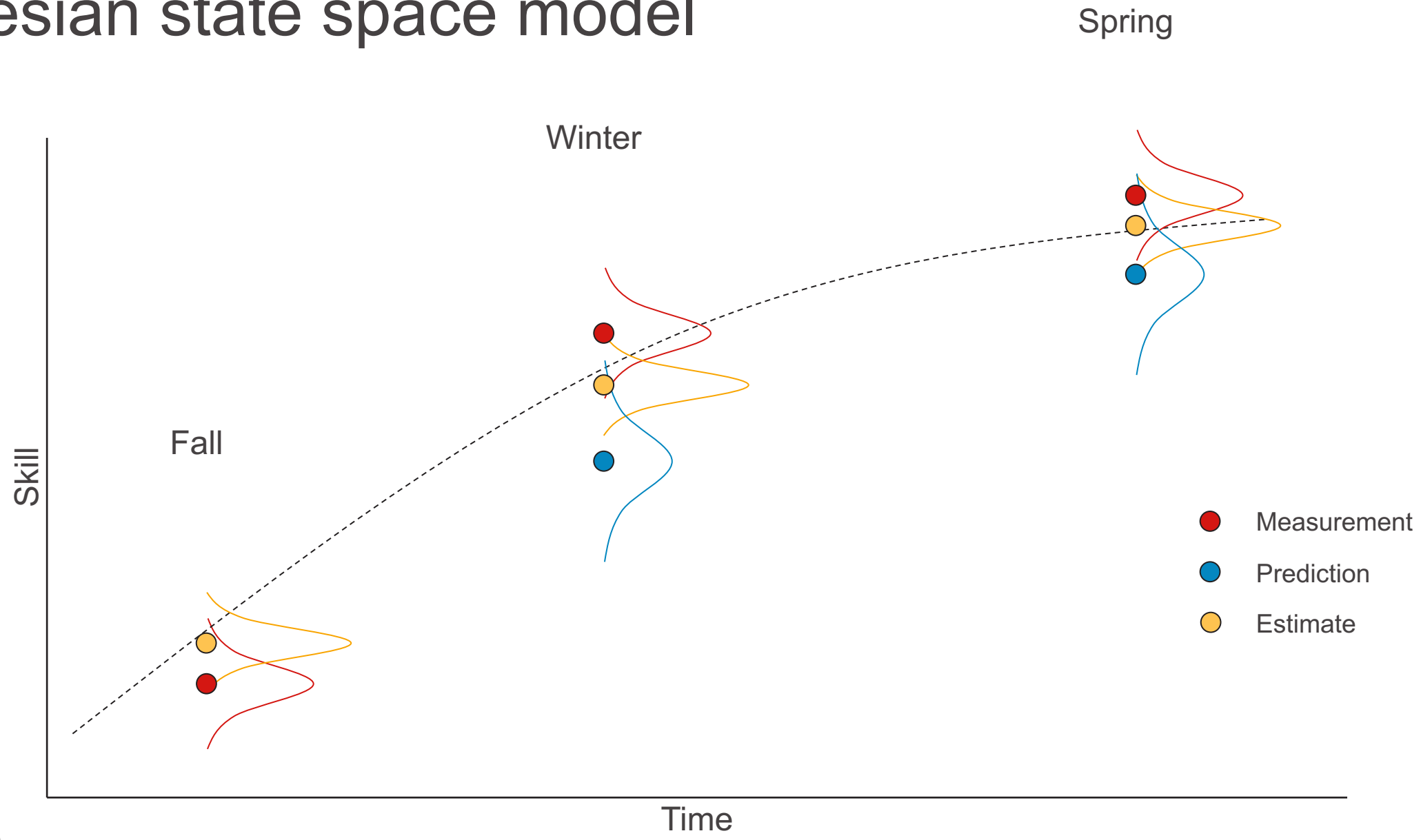
Here, K_n is the "Kalman gain" – governs weight given to prediction vs. measurement. If assumptions met, gives minimum possible RMSE.



nwea

Methods

Bayesian state space model



nwea

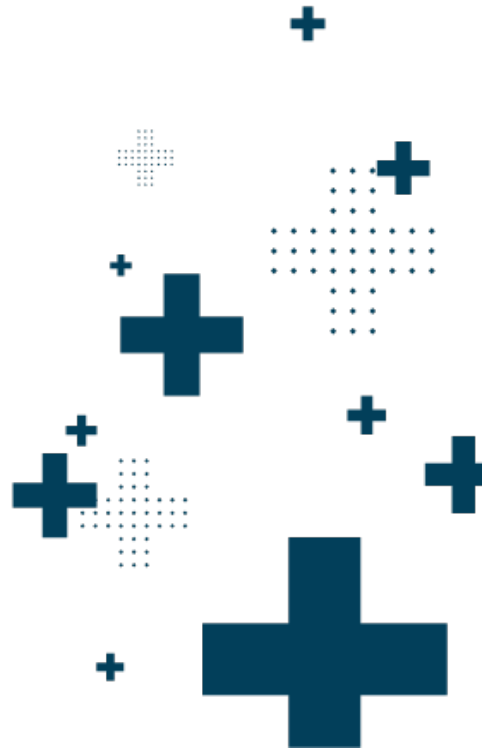
Bayesian state space model

Based on Kalman filter but uses posterior sampling to jointly estimate student latent variables at each time step

- \+ Relax need for known process noise
  - – Important since latent trait (student parameter) is unobservable
- \+ Estimates latent student parameters a_t jointly
- \+ Relax need for Gaussian posterior
- \+ Supports any arbitrary expected growth model
- \+ Allows for generating smoothed scores for all 3 tests

$$\theta_{i,t} \sim \mathcal{N}\left(a_{i,t}, \mathrm{SEM}_\theta\right)$$

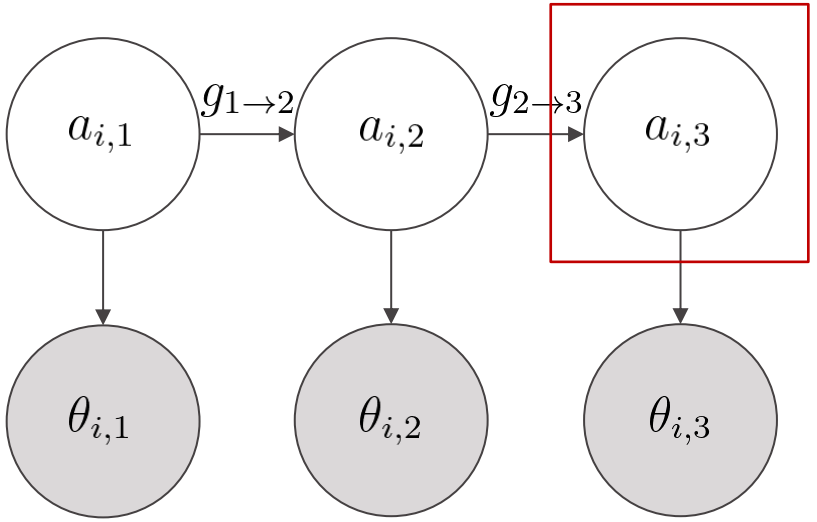
$$a_{i,t} \sim \mathcal{N}\left(a_{i,t-1} + g(a_{i,t-1}, \Delta t), \gamma\right)$$



nwea

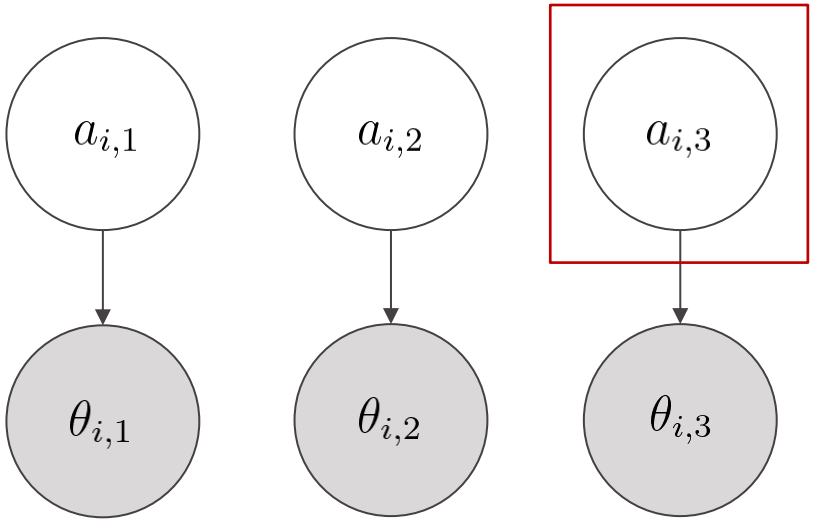
Bayesian state space model

State space model (SSM)



Baseline model

+ Estimates latent variable independently at each time step



nwea

Data

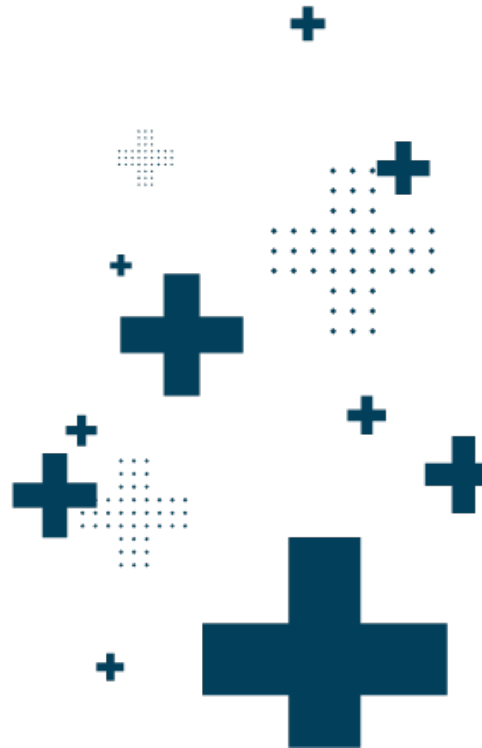
Reading & math computer-adaptive tests for 5th grade cohort, 2018-2019 school year

Real dataset selection:

- School cohorts of size 51-100, 101-150, …, 201-250 (10 each, 40 cohorts total)
  - 1,976 in reading & 2,009 students in math
  - Very low SEM (avg. 0.15 math and 0.17 reading), rather long tests (avg. 53 items math and 40 items reading)

Simulated data:

- Simulated using real data as a basis for test 1 (fall) scores, then projected out using our conditional growth norms model
- 100 replications of each test
- 9 datasets total: 3 levels of process noise × 3 levels of measurement noise

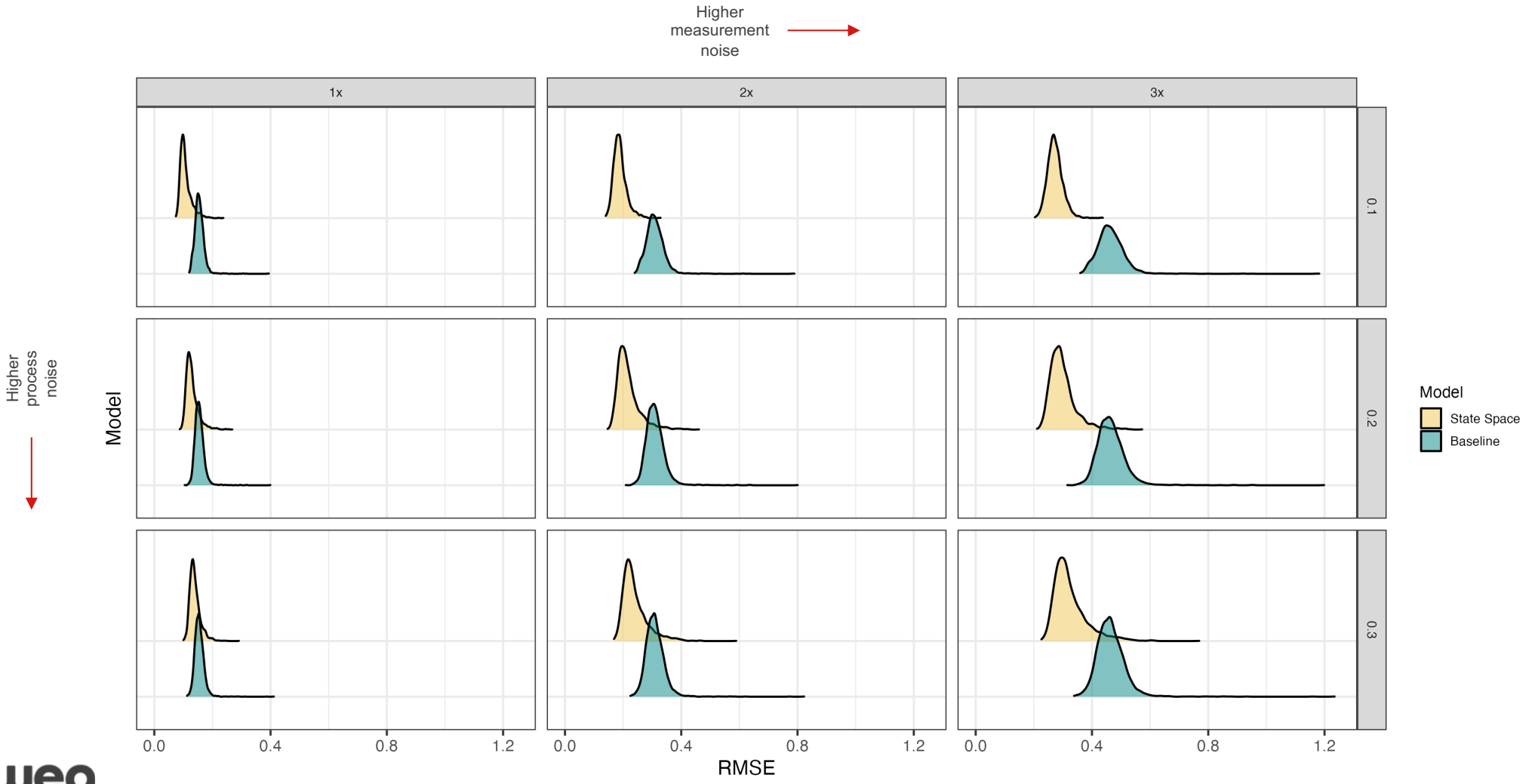


nwea

Results

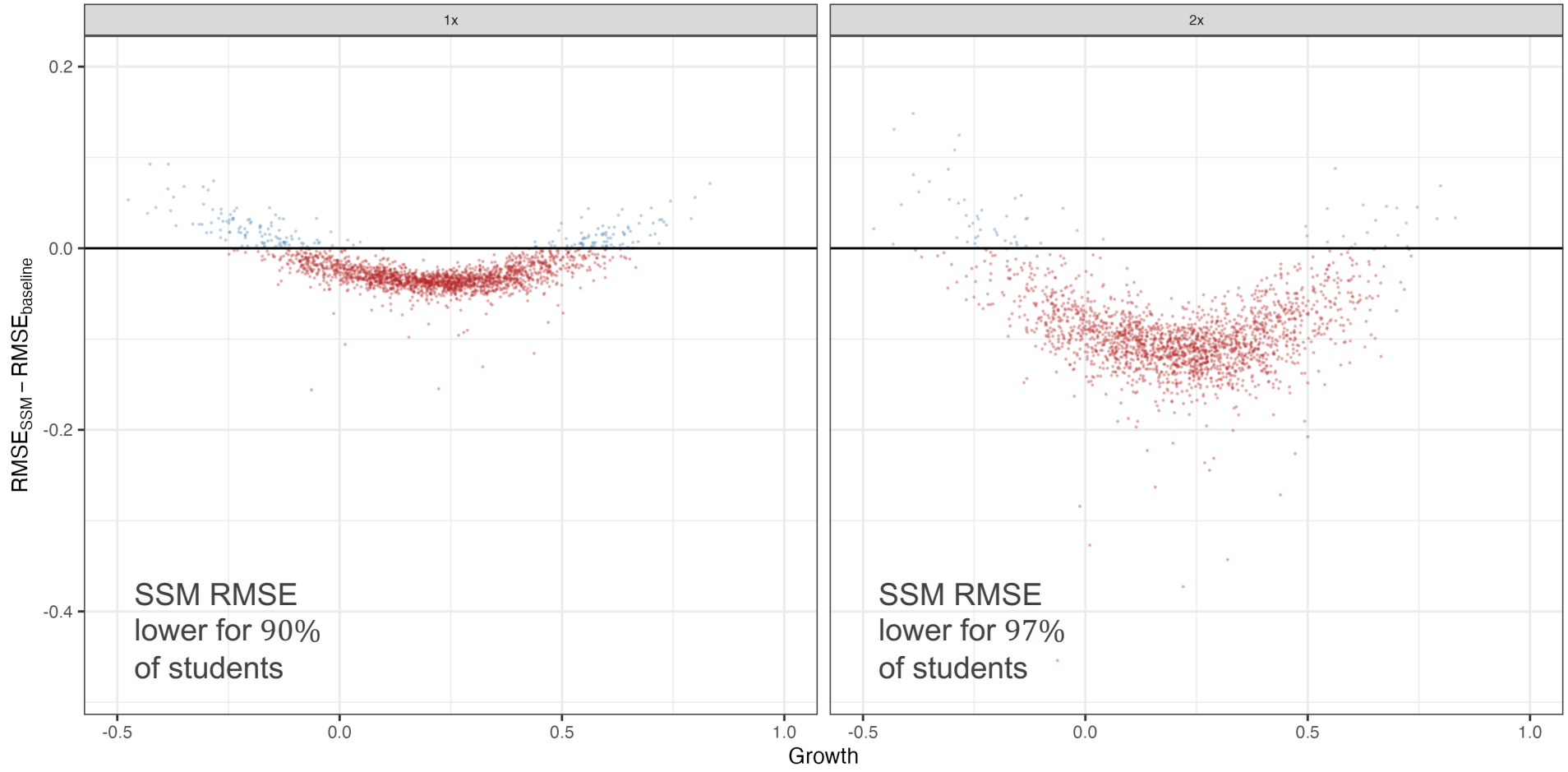
Parameter recovery

RMSE (taken over replications, within students)



Parameter recovery

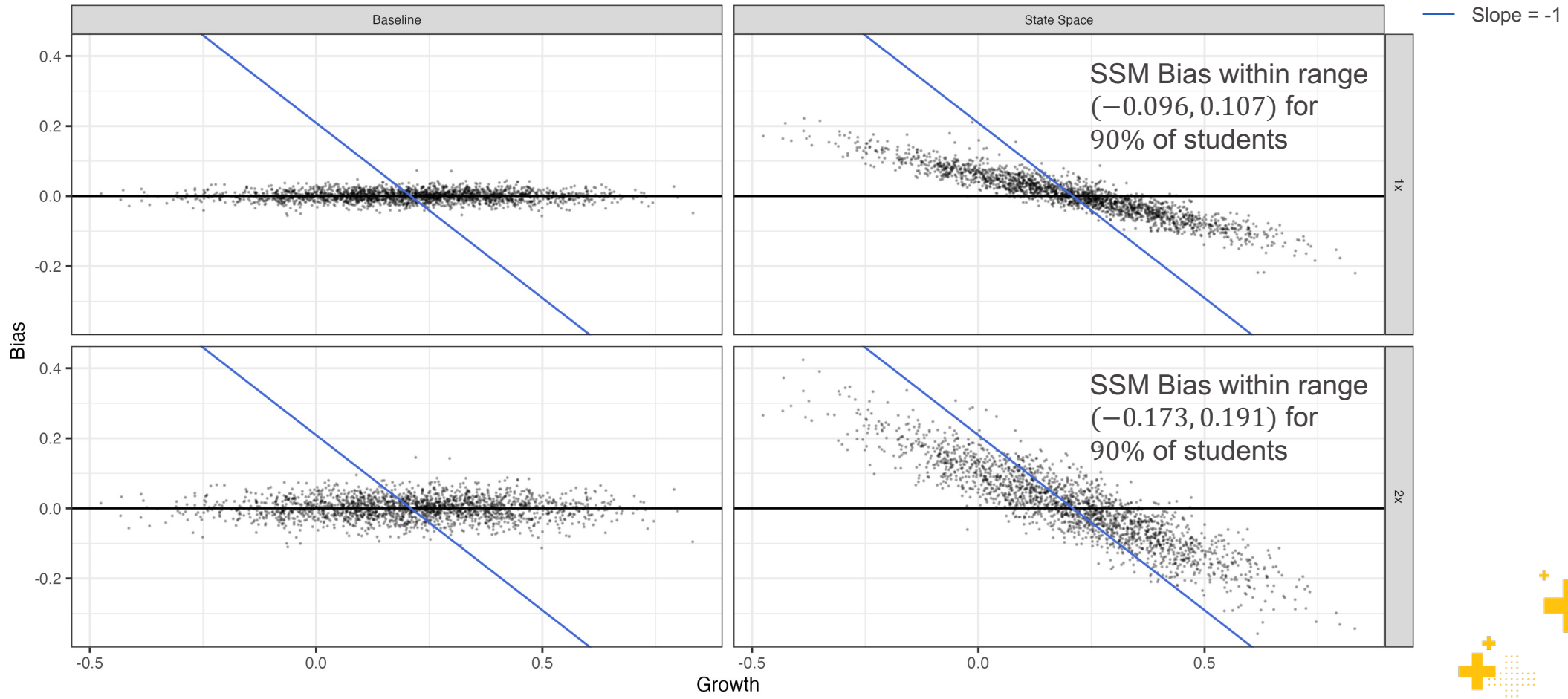
RMSE (taken over replications, within students)



nwea

Parameter recovery (cont'd)

Bias (averaged over replications, within students)



nwea

Conclusion

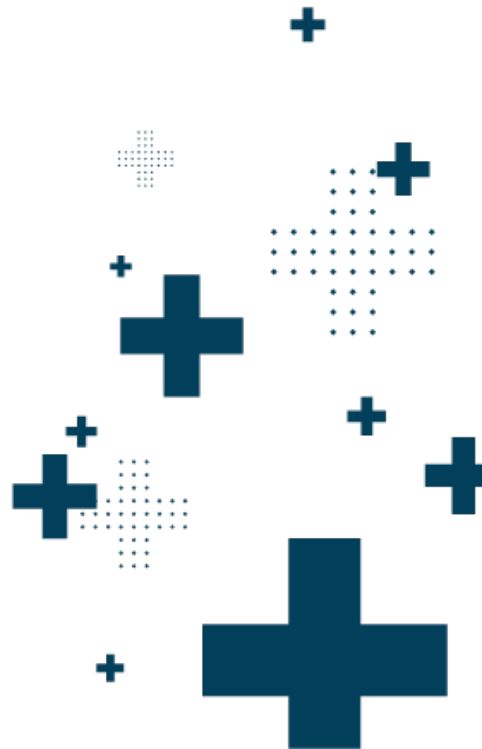
Conclusion

Goal: Combine 3 scores into a summative score that yields a good estimate of latent trait at the final time point. Our method was successful on simulated data for vast majority of students

Nice-to-have: Generate smoothed scores for all tests

Findings

+ Simulated data
  - SSM performance better than baseline in terms of RMSE, at the expense of introducing some bias
+ Real data (not shown in results)
  - Evidence for better model-data fit for SSM than baseline

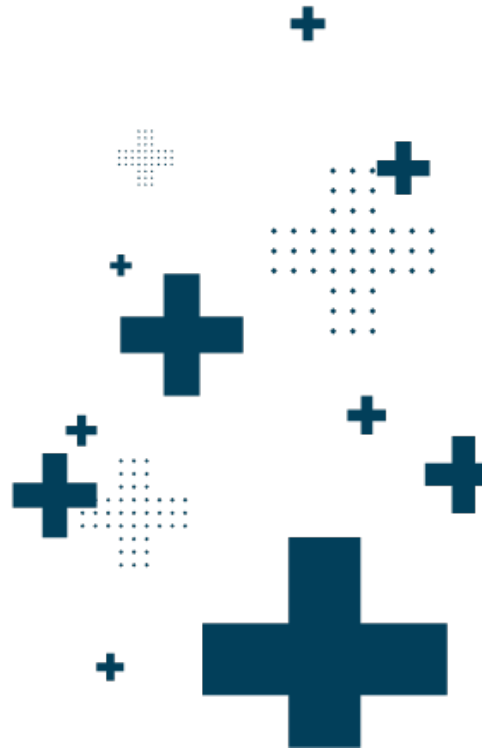


nwea

Conclusion

Limitations and future work

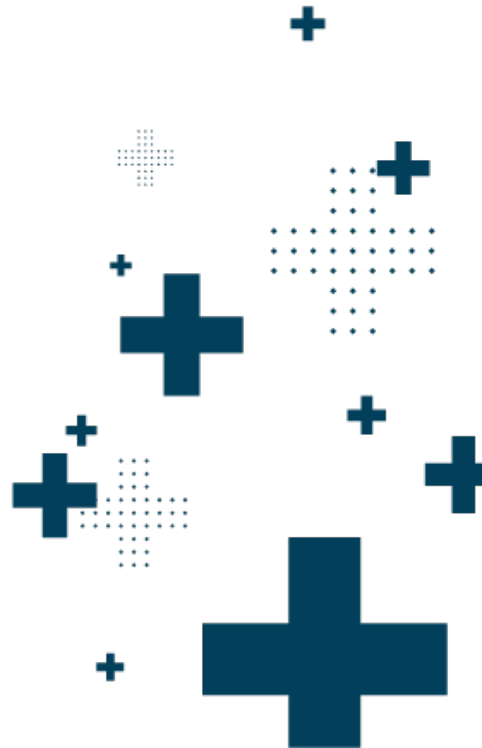
- More sophisticated method for estimating process noise?
- Assumes construct continuity between tests. Can model be adapted for non-continuous content across assessments?
- Requires growth model. Could this be learned jointly with the SSM?
- Can SSM estimation be used to reduce test length or improve computer-adaptive tests (CATs)?



nwea

References

- L. L. Wise, “Picking up the pieces: Aggregating results from through-course assessments,” 2011.
- A. V. Moere and S. Hanlon, “A bayesian approach to improving measurement precision over multiple test occasions,” Language Testing, 2020.
- R. Zwick and R. J. Mislevy, “Scaling and linking through-course summative assessments,” The Computer Journal, 2011.



nwea

Thank you

nwea